ASSOCIATIONS OF UNCONVENTIONAL NATURAL GAS DEVELOPMENT WITH HEART FAILURE HOSPITALIZATION AND B-TYPE NATRIURETIC PEPTIDE AND EFFECT MODIFICATION BY HEART FAILURE PHENOTYPE

by

Tara Patricia McAlexander

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
July, 2019

Abstract

Persons with heart failure are at risk for frequent hospitalization, diminished quality of life, reduced life expectancy, and are susceptible to environmental and contextual effects related to unconventional natural gas development (UNGD) activity. We assessed whether four UNGD activity metrics by phase were separately associated with hospitalization for heart failure or B-natriuretic peptide (BNP) levels in blood among subjects with heart failure, and we evaluated effect modification by heart failure with preserved (HF*p*EF) or reduced (HF*r*EF) ejection fraction on these associations. We obtained electronic health records for subjects with heart failure who were seen at a Geisinger facility in Pennsylvania between 2008-2015. We utilized multilevel logistic regression and generalized estimating equations (GEE) to evaluate associations of 30-day metrics of UNGD activity with odds of hospitalization for heart failure and odds of BNP concentration ≥ 400 pg/mL. Comparing subjects in the fourth to the first quartile of UNGD activity, adjusted associations (OR [95% confidence interval]) with hospitalization were 1.70 (1.35–2.13), 1.80 (1.35–2.40), and 1.62 (1.07–2.45) for the pad preparation, well stimulation, and production metrics, respectively; each metric evidenced exposure-effect relations. Only the production metric was associated (OR [95% confidence interval]) with BNP and evidenced exposure-effect relations of increasing magnitude of association across quartiles, of 1.36 (1.08–1.71), 1.42 (1.05–1.93), and 1.52 (1.07–2.17) for the second, third, and fourth (vs. first) quartiles. In the effect modification analysis of hospitalization, we did not find evidence in favor of our hypothesized moderation, but did observe stronger associations of UNGD activity metrics with heart failure hospitalization among those who were phenotyped (subjects with HF*p*EF and HF*r*EF) vs. those who could not be phenotyped (p for global significance = 0.03 for pad; 0.009 for spud; 0.4 for stimulation; and 0.08 for production metrics). In the BNP analysis, there was a strong

independent effect of HF*r*EF phenotype with greater BNP concentrations and some evidence of effect modification of UNGD and BNP associations by HF*r*EF phenotype. However, these associations were difficult to interpret given overlapping confidence intervals. These results suggest that UNGD activity could exacerbate existing conditions in exposed heart failure subjects.

Thesis readers: Drs. Brian Schwartz, Karen Bandeen-Roche, Jessie Buckley, and Erin Michos
Alternates*:* Drs. Brian Caffo, Ghassan Hamra, and Genee Smith

Acknowledgements

This work would not have been possible without a number of individuals who provided support, directly or indirectly, to this research.

First, this work could not have been possible without the mentorship of Brian Schwartz, who has been a wonderful mentor and thesis advisor. Thank you for always encouraging me to ask the tough questions and for dedicating so much of your time to guiding me through this dissertation process. I am a stronger, more careful scientist because of your mentorship, and I have learned a great deal from working together.

Second, I have been fortunate to have the honor of working with a broader committee of experts who have been kind and generous with their expertise and advice. Thank you to Karen Bandeen-Roche for helping with all of this work's statistical challenges. Thank you to Jessie Buckley for supporting my growth as an environmental epidemiologist, within the context of this research as well as throughout my coursework. This work has been strengthened by the input and expertise of two cardiologists: Erin Michos and Bill McEvoy, who have shared their time and energy to support this research.

Third, I am grateful to the Johns Hopkins Bloomberg School of Public Health and the Department of Environmental Health and Engineering for the support and training I have received. I am confident that my time here has made me a better, more rigorous environmental epidemiologist.

I would like to thank my family – Mom, Dad, Laura, Rick, Dan, PernaLyn, Danny, Ruby, and Margot – who have supported me throughout all my endeavors, encouraged my growth, and have filled my life with so much joy. Lastly, I owe many thanks and gratitude to Ryan Nagle, for being a supportive partner, always.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Equations

Chapter 1: Introduction

1.0 RATIONALE

This dissertation work was borne from increasingly convincing environmental and epidemiologic evidence indicative of the negative environmental, societal, and health impacts of unconventional natural gas development (UNGD). UNGD has been colloquially referred to as "fracking," but this term was designed by the industry's opponents to have negative connotations, only refers to a brief phase of the entirety of UNGD, and will not be used herein. UNGD involves obtaining natural gas from shale resources below the earth's surface and consequently has resulted in areas of concentrated development and activity, with adverse environmental (i.e., air pollution, groundwater contamination, noise) [1-5] and contextual (community changes, crime, and traffic) impacts [6-8]. UNGD has been associated with a number of adverse health outcomes (low birth weight, small for gestational age [9, 10]; preterm birth [11-13]; congenital defects [14]; three types of asthma exacerbations [15, 16]; migraine, fatigue, nasal and sinus symptoms [17]; and depression symptoms [18]) in environmental epidemiologic studies, however all of these studies have evaluated outcomes that concern early and mid-life populations; none have evaluated biological markers informing candidate exposure pathways, and none have evaluated any biologically important effect modification.

To date, no environmental epidemiology studies have evaluated UNGD activity in relation to heart failure, a prevalent, severe, and costly disease that typically affects older individuals [19-21]. Individuals with heart failure are prone to frequent hospitalization, which can be exacerbated by air pollution [22-26] and other community impacts of UNGD [27]. This research therefore evaluated associations between UNGD activity and hospitalization for heart failure. Further, this dissertation evaluated

associations between UNGD activity and both inpatient and outpatient laboratory measures of B-type natriuretic peptide (BNP), a clinical biomarker of heart failure diagnosis and prognosis [28-32]. We also evaluated effect modification of these associations by heart failure phenotypes (i.e., heart failure with preserved [HF*p*EF] or reduced [HF*r*EF] ejection fraction), two distinct presentations of heart failure with differing pathophysiology, which allowed us to understand biologic mechanisms underlying associations between UNGD activity and hospitalization and between UNGD activity and BNP levels. Understanding the associations between UNGD activity and heart failure would fill a gap in the epidemiologic literature regarding potential health impacts of UNGD in older populations, whom we would expect to be most susceptible to the environmental and community impacts associated with UNGD. Further, this work advanced the environmental epidemiology of heart failure, particularly with respect to heart failure phenotypes and understanding differential environmental associations with BNP.

1.1 UNGD: AN OVERVIEW

Unconventional natural gas development (UNGD) has increased rapidly in the United States in the past ten years and is expected to continue to grow [33, 34]. In contrast to conventional natural gas, which is procured from geological resources that do not require advanced technologies to obtain, unconventional natural gas requires the tapping of shale resources that can sit at least a mile below the ground's surface [35] and involves hydraulic fracturing and/or horizontal drilling. UNGD wells are typically between 5000-10,000 feet deep and can extend up to two horizontal miles from the drill site [36]. The UNGD process involves, in order, well pad preparation (e.g., clearing of land), well drilling (both vertical and horizontal), well stimulation (i.e., hydraulic fracturing), gas extraction, storage and transport. A distinguishing feature of UNGD is

well stimulation, which involves the injection of fracking fluid (a mixture of water, chemical lubricants, and proppants [e.g., sand or ceramic materials to keep fractures open]) into the well to force shale rock to fracture, releasing natural gas into the well for extraction. Consequently, UNGD wells require the use and transport of substantive equipment, and compared to conventional natural gas development, UNGD produces vastly more natural gas and waste byproducts (e.g., wastewater, drilling lubricants, sand, and radionuclides) [35, 37].

The process of UNGD has been associated with adverse environmental impacts that are relevant to public health. First, UNGD has been associated with increases in air pollutants (particulate matter less than 2.5 microns in diameter [$PM_{2.5}$], volatile organic compounds [VOCs], oxides of nitrogen [NO_x], oxides of sulfur [SO_x], and ozone [O_3]) through a variety of sources and phases within the UNGD process [38-41]. Some of these air pollutants are regional (e.g., $PM_{2.5}$, NO_x, O_3) [39], whereas VOCs are more local and are characterized by episodic release into the environment from the UNGD process [41]. Phases of UNGD activity that have been specifically linked to these emissions include drilling, hydraulic fracturing, production, compressor stations, and transport [39, 40]. In addition to the direct release of air pollutants into the environment, air pollutants can vaporize from injected fracking fluid, subsurface water, and storage pits and tanks used to store “flow back” wastewater, which is water that is recovered from wells after being used to hydraulically fracture shale [37, 42-44]. Because of the chemicals used in the hydraulic fracturing process, and because of natural gas (methane) itself is a water contaminant, groundwater and drinking water contamination near UNGD activity has been documented [2, 5, 45], although the extent of groundwater and drinking water contamination is highly variable due to geologic factors [46, 47]. Additionally, due to the necessity of transporting equipment, personnel, and natural gas, UNGD operations have been associated with increased truck traffic and traffic accidents [6, 48]. Traffic, and

truck traffic in particular, has been associated with increased noise and higher air pollution levels [22-24, 49], suggesting that there could be secondary environmental impacts from the volume of traffic associated with UNGD. **Figure 1.1** provides a visual representation of how negative environmental impacts can occur as a result of unconventional natural gas development at various stages.

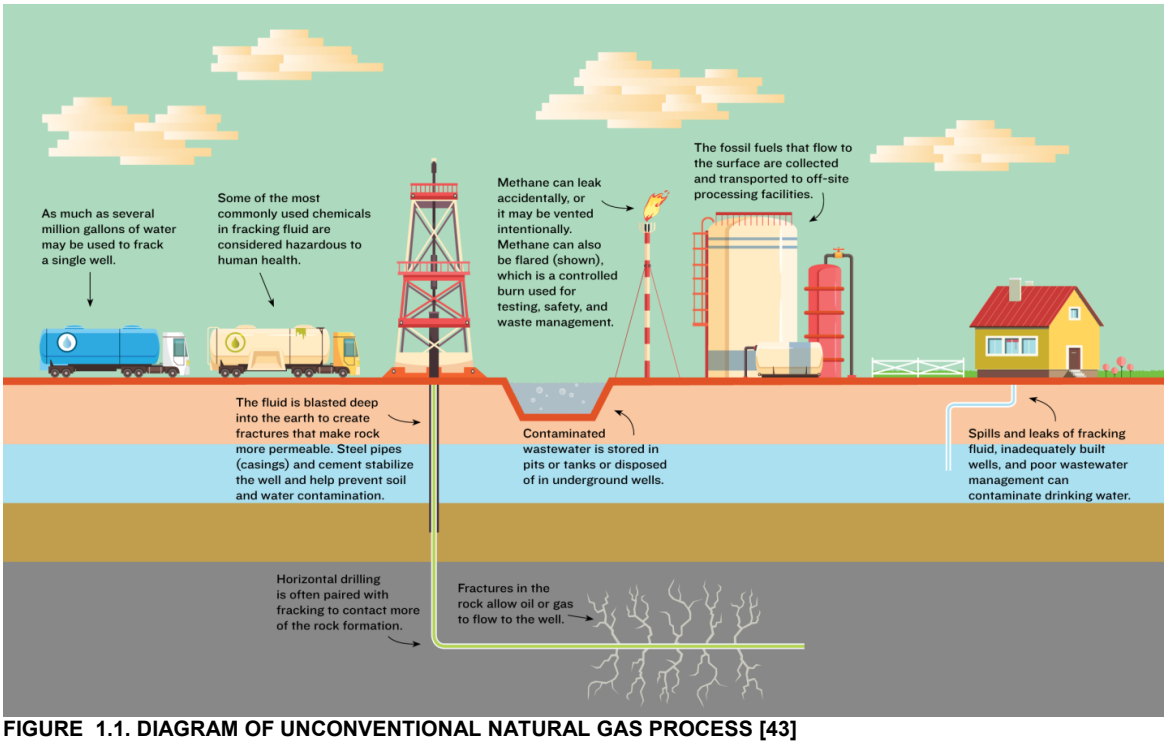


FIGURE 1.1. DIAGRAM OF UNCONVENTIONAL NATURAL GAS PROCESS [43]

Adding to the intensity of its environmental impacts, horizontal drilling allows for multiple wells on one well pad, meaning that a lot of natural gas is extracted from a very small area, and thus activities are very concentrated [22, 27]. This does not, however, negate the fact that UNGD results in the loss of natural vegetation and undeveloped land; a 2010 study by the Nature Conservancy estimated that UNGD requires the development of, on average, 8.8 acres of forest land for well pad development and the building of associated roads and infrastructure [34], which has relevance to heart failure because greenness has been associated with improved cardiovascular outcomes [50-52]. In addition to the loss of natural habitat, the influx of necessary personnel, machinery, and

equipment for UNGD results in secondary and contextual impacts, and several studies have documented increases in community stressors, such as crime and traffic, due to UNGD [9, 22, 53, 54]. Additionally, UNGD can have negative impacts on housing and property values in areas of heavy development [7, 55]. Compared to conventional natural gas development, UNGD may pose additional stress risks stemming from these secondary impacts and community changes [27, 54].

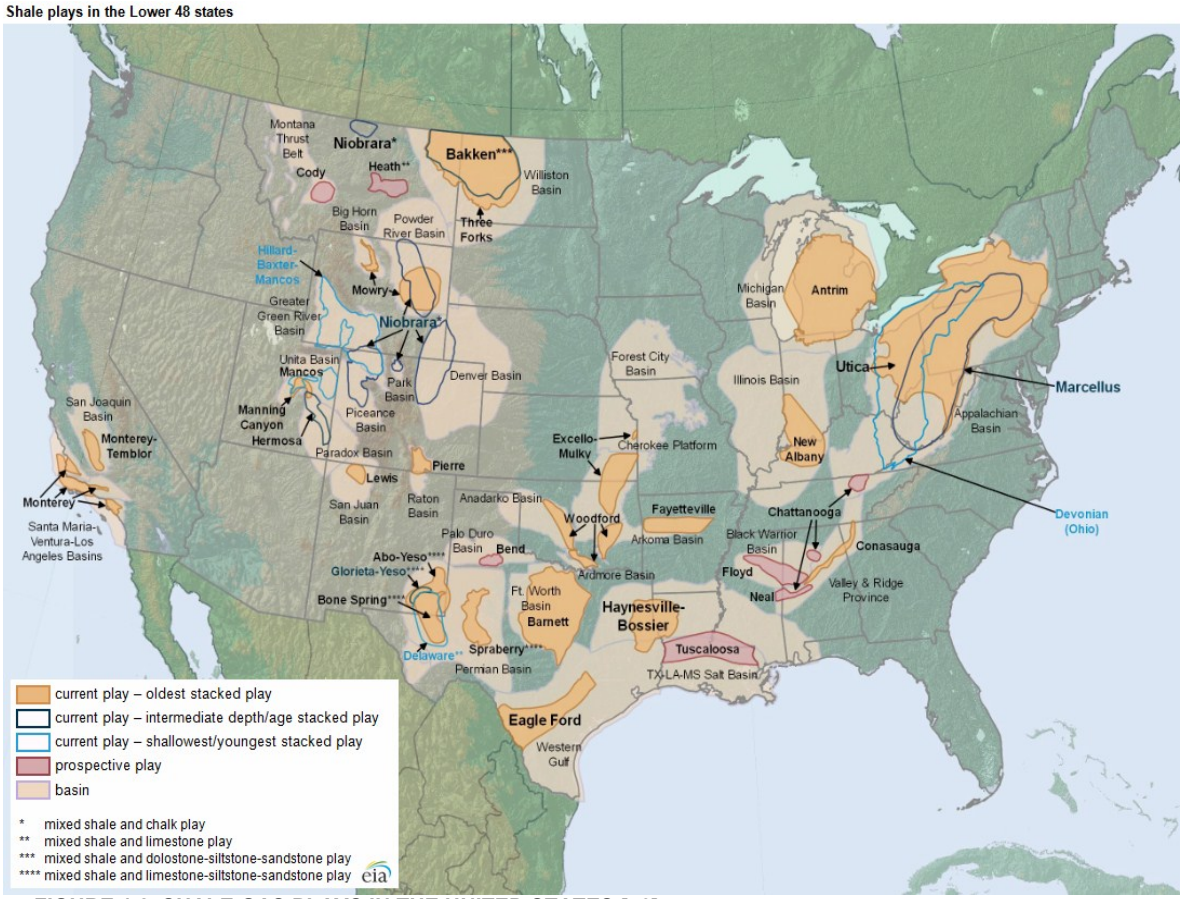


FIGURE 1.2. SHALE GAS PLAYS IN THE UNITED STATES [56]

1.1.2 UNGD IN PENNSYLVANIA

Shale gas deposits are located throughout the US (**Figure 1.2**), and the UNGD industry has been established in at least 16 states [57, 58]. In particular, shale gas resources in Colorado, Texas and Pennsylvania have seen intensive development, eliciting mixed responses from communities and environmental advocates [59, 60]. Of

these states, Pennsylvania has experienced some of the most rapid development where over 9000 wells have been drilled in the Marcellus shale since 2004, but mainly since

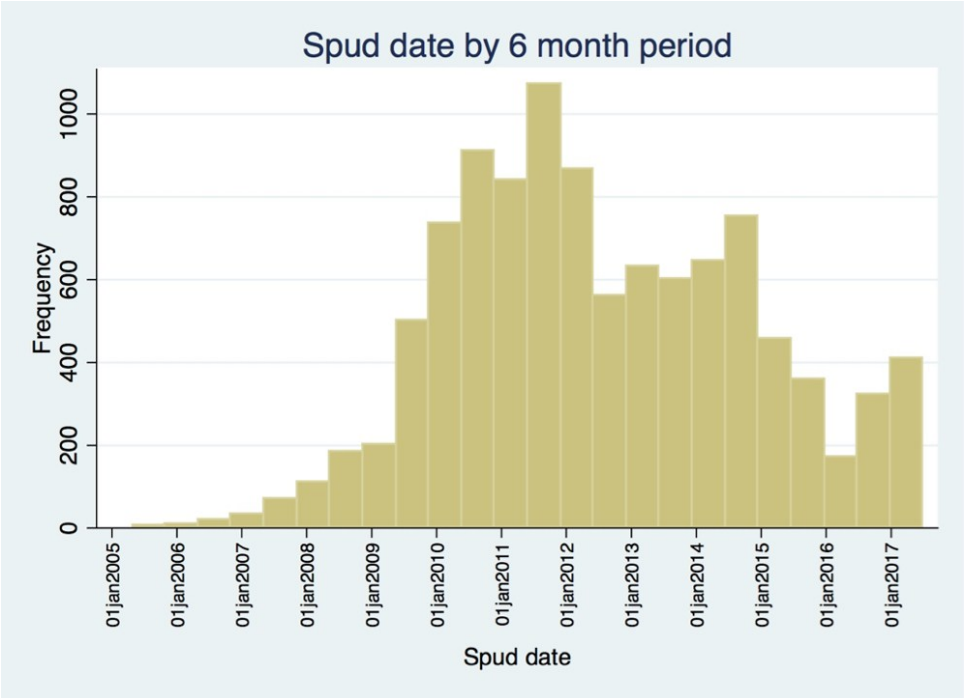


FIGURE 1.3. TIME SERIES OF NEWLY SPUDDED (DRILLED) NATURAL GAS WELLS IN PENNSYLVANIA

2009, according to regulatory data (**Figure 1.3**) [61], making Pennsylvania the second-largest natural gas producing state in the US [43].

1.1.3 UNGD AND HEALTH: EPIDEMIOLOGICAL EVIDENCE AND EXPOSURE PATHWAYS

Epidemiologic studies to date have attempted to characterize and quantify UNGD activity through the use of UNGD activity metrics and [9, 12, 15, 62]. Growing epidemiologic evidence supports associations between UNGD activity metrics and population health impacts (e.g., pregnancy outcomes [low birth weight, small for gestational age, preterm birth, congenital defects], asthma exacerbations, symptoms of chronic rhinosinusitis, migraine headaches, fatigue, self-reported stress, and depression symptoms [9, 13, 17, 18, 57, 63]). An advantage of the use of UNGD activity metrics in epidemiologic studies is that they likely reflect multiple exposure pathways, including the

contextual effects of UNGD activity. UNGD is suspected to exacerbate stress through its secondary impacts (i.e., land use changes, increases in traffic, and negative impacts on property values [27, 55], so GIS-based UNGD activity metrics will likely reflect these secondary impacts as well. A disadvantage to using UNGD activity metrics in epidemiologic studies, however, is that it is not known whether the associations between UNGD activity metrics and health outcomes observed in epidemiologic studies are due to specific chemical (e.g., air pollution) or physical (i.e., noise, light, vibration) agents, or to stress-related exposure pathways [1, 64], each of which should be captured by the GIS-based UNGD activity metrics employed to date. Despite these limitations, a growing body of epidemiologic evidence has found associations between UNGD activity and biologically plausible adverse health outcomes. No prior epidemiological studies have isolated specific exposure pathways underlying observed associations between UNGD activity metrics and any health outcomes [9, 11, 13, 15, 65]. While we did not address this limitation of previous studies through improved exposure assessment, we utilized UNGD activity metrics to evaluate associations between UNGD and heart failure outcomes. Importantly, we had retrospective health data, so retrospective assessment of UNGD activity was necessary, and therefore UNGD activity metrics were helpful.

To date, no epidemiologic studies have evaluated UNGD activity in relation to heart failure outcomes, a common and severe end-stage cardiovascular disease. However, rationale for epidemiologic studies of UNGD and heart failure are supported by other epidemiologic studies of UNGD and health [66]. A 2018 study in Colorado, for example, has linked oil and gas activity to slightly higher levels of plasma concentrations of interleukin-1β (IL-1β) and tumor necrosis factor alpha (TNF-α), two inflammatory markers [66]. Although the sample size of this study included only 97 individuals living near oil and gas activity (i.e., not only UNGD activity), both of these inflammatory markers have been specifically implicated in endothelial dysfunction and the progression

of heart disease [67-70] in older adults [71, 72], which lends support to the biologic rationale for studying heart failure in relation to UNGD.

1.2 Epidemiology of Heart Failure

1.2.1 Overview

Heart failure refers to a condition in which the heart is unable to pump blood effectively, causing a cascade of symptoms that impact patient quality of life, morbidity, and mortality. The Centers for Disease Control and Prevention (CDC) estimate that 5.7 million Americans have heart failure [73]. Prevalence of heart failure is projected to increase by 46% from 2012 to 2030; at this rate, there will be more than 8 million Americans with heart failure by the year 2030 [74, 75]. Risk factors for heart failure include classic cardiovascular disease risk factors such as coronary artery disease, high blood pressure, diabetes, tobacco smoking, lack of physical activity, poor diet, and obesity [73]. Although not all heart failure cases result in immediate death, it is estimated that in 2009, 1 in 9 deaths in the United States were in some way due to heart failure (i.e., heart failure was mentioned on the death certificate); the average duration of survival after diagnosis is only five years, and 10-year survival rates are estimated to be as low as 10% [76, 77].

Symptoms of heart failure include shortness of breath, fatigue, and swelling of the legs due to fluid buildup (edema) [75]. Because heart failure can make physical activity difficult and results in a number of physical symptoms, living with heart failure can have negative impacts on quality of life, and mental health comorbidities such as depression often accompany the condition [78-80]. Furthermore, the disease is costly; CDC estimates that heart failure results in over $30 billion in direct and indirect health care costs, including missed work days, medication costs, and health care services,

annually [21, 73]. By 2030, costs of heart failure treatment are projected to increase to $53 billion (direct) and $70 billion (total) [74].

1.2.2 Heart failure physiology, phenotypes, and the heart failure spectrum

Traditionally, heart failure includes two phenotypes: heart failure with preserved ejection fraction (HF*p*EF) and heart failure with reduced ejection fraction (HF*r*EF) [81]. The ejection fraction is a measure of the proportion of blood that the left ventricle is able to effectively pump out of the heart relative to the quantity that is pumped into the left ventricle from the left atrium. In HF*r*EF or systolic heart failure, the heart muscles cannot contract to efficiently pump blood, and thus the total volume of blood that is pumped out of the heart is substantially reduced [82]. In HF*p*EF or diastolic heart failure, the heart muscle stiffens or does not adequately fill with blood, but the relative volume of blood pumped out of the heart remains preserved [82]. Presentation of symptoms is similar for HF*p*EF and HF*r*EF (i.e., shortness of breath, fatigue, fluid retention), and each accounts for roughly 50% of heart failure cases [83]. Epidemiological literature has found common risk factors among phenotypes for both developing heart failure and for heart failure hospitalization, including age, race/ethnicity, sex, obesity, tobacco smoking, physical inactivity, diabetes, stress, hypertension, and previous myocardial infarction and coronary artery disease diagnoses [73, 77, 81, 84].

However, there is evidence that some risk factors differ between HF*p*EF and HF*r*EF, which has implications for epidemiologic study [81, 82, 85]. Studies of the precipitating factors in HF hospitalization, for example, have found that patients with HF*p*EF were more likely to be female, obese, have poor blood pressure control, lower hemoglobin levels, higher exposures to NO_2, have had a previous episode of atrial fibrillation, and more commonly had respiratory infections than patients who were admitted for HF*r*EF [86-88]. There is also evidence that the prevalence of HF*p*EF is

increasing relative to HF*r*EF, possibly due to increases in HF*p*EF-associated comorbidities such as obesity and diabetes [77, 89-91].

In recent years, there has been evolved thinking regarding the pathology of heart failure which focuses more on the intracardiac (e.g., cardiomyopathy, hypertrophy, ventricular remodeling) and extracardiac abnormalities (e.g., diabetes, obesity, hypertension) shared between HF*p*EF and HF*r*EF patients, rather than on the differences between the two phenotype groups [92-96]. Some researchers have even argued that the two phenotypes should be viewed less as distinctive conditions but rather as differing, yet overlapping, presentations of the same disease along a spectrum [92, 97]. For these reasons, we chose to evaluate heart failure outcomes with differentiation between HF*p*EF and HF*r*EF subjects.

1.3 B-TYPE NATRIURETIC PEPTIDE (BNP)

Natriuretic peptides are a family of vasoactive neuroendocrine proteins and include atrial natriuretic peptide (ANP), B-type natriuretic peptide (BNP), and C-type natriuretic peptide (CNP) [98, 99]. ANP is expressed in the atria, ventricles, kidney, adipose tissue, and in the brain; BNP is primarily expressed by the ventricles, but also by atria and the brain; lastly, CNP can be expressed in the bone, brain, endothelial tissues, and throughout the heart [98]. In recent years, BNP in particular, has shown promise for the diagnosis and management of heart failure [28, 100, 101]. Epidemiologic studies increasingly also measure NT-pro-BNP, which is a biologically inert byproduct of BNP synthesis but has a longer half-life, and find similar clinical interpretations as do studies of BNP [100-103]. Serum concentrations of both BNP and NT-pro-BNP are measured at Geisinger laboratories by immunoassay [104].

1.3.1 BNP SYNTHESIS

The most common source of BNP synthesis is from the ventricles in response to intracardiac pressures, such as volume overload which occurs in heart failure patients [98, 105, 106]. Under these pressures, *de novo* synthesis of BNP occurs in the ventricular myocardium due to rapid gene expression [105]. BNP is synthesized as a 108-amino acid prohormone (proBNP), and, when secreted, is cleaved into two byproducts: NT-proBNP (biologically inert, 76 amino acids) and BNP (biologically active, 32 amino acids) [105, 107]. **Figure 1.4** provides a visual schematic of BNP synthesis and secretion into systemic circulation from the ventricular myocardium. Once in systemic circulation, BNP is cleared by receptor mediated mechanisms (e.g., natriuretic peptide receptor-A [NRP-A], natriuretic peptide receptor-C [NRP-C]) and excreted by the kidneys [98, 107, 108]. Of note, the gene that codes for BNP synthesis, *Nppb*, has been identified in humans on chromosome 1 at 1p36.2 [98, 109]. Several single nucleotide polymorphisms (SNPs) have been identified in the *Nppb* gene, although these SNPs seem to only affect proBNP, but not the circulating 32 amino acid form of BNP which is found in circulation [98]. This implies that BNP might be more appropriate than NT-proBNP in epidemiologic studies, since there could be more inter-individual variability in measures of NT-proBNP than for BNP.

1.3.2 BNP BIOLOGY

Circulating BNP has a number of biologic functions that include decreasing blood pressure, increasing sodium excretion and water excretion (natriuresis and diuresis); regulating sympathetic nervous system activity; inhibiting the renin-angiotensin-aldosterone system; modulating vasopressin, endothelin, and cytokine activity, and involvement in cardiac remodeling [110, 111]. Interestingly, BNP seems to be a natural defense mechanism in response to high intracardiac pressures; its circulation is involved

in neuroendocrine signaling between the heart, kidneys, and brain to maintain homeostasis (**Figure 1.5**) [111].

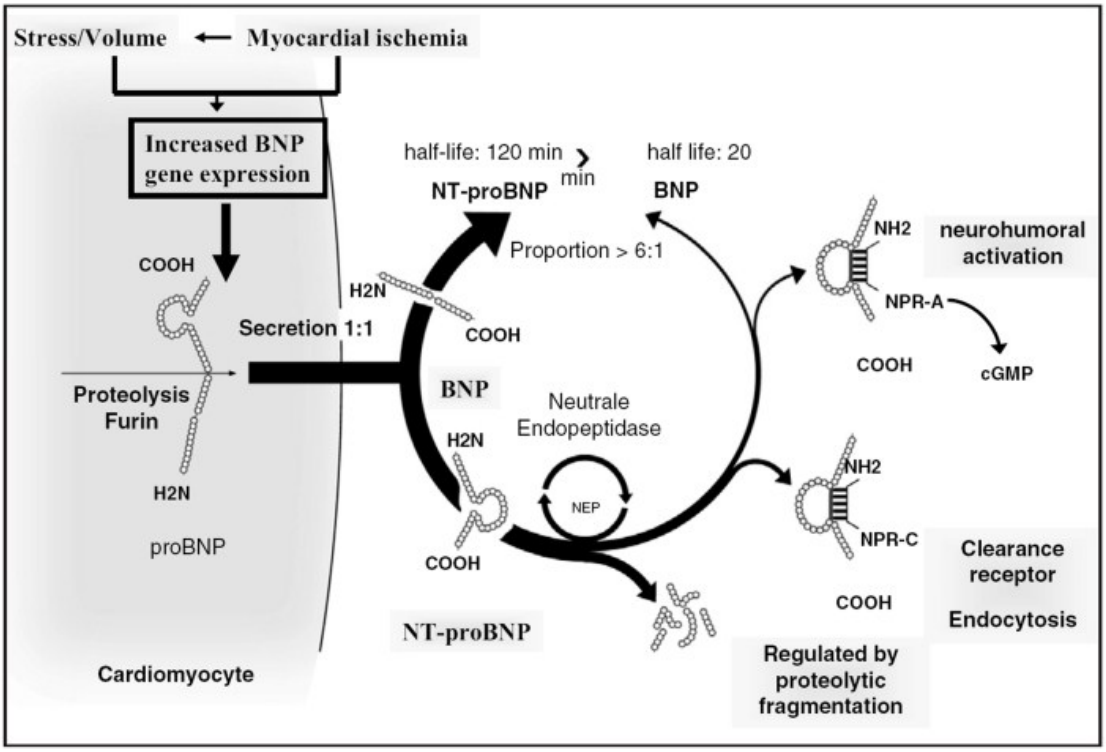


FIGURE 1.4. SCHEMATIC OF BNP SYNTHESIS [105]

There is some evidence that BNP can interface with the HPA axis [112, 113], although the mechanisms that link BNP with HPA and psychosocial stress are not very well understood. Some studies show decreases in BNP in relation to psychosocial stress [114], while others have illustrated that psychosocial stress leads to increases in BNP [115], and others have shown that BNP increases were associated with greater reports of emotional anger [116]. Yet, as BNP was first isolated from a porcine brain in 1988 [117], the complex biochemistry of BNP is not entirely understood [118].

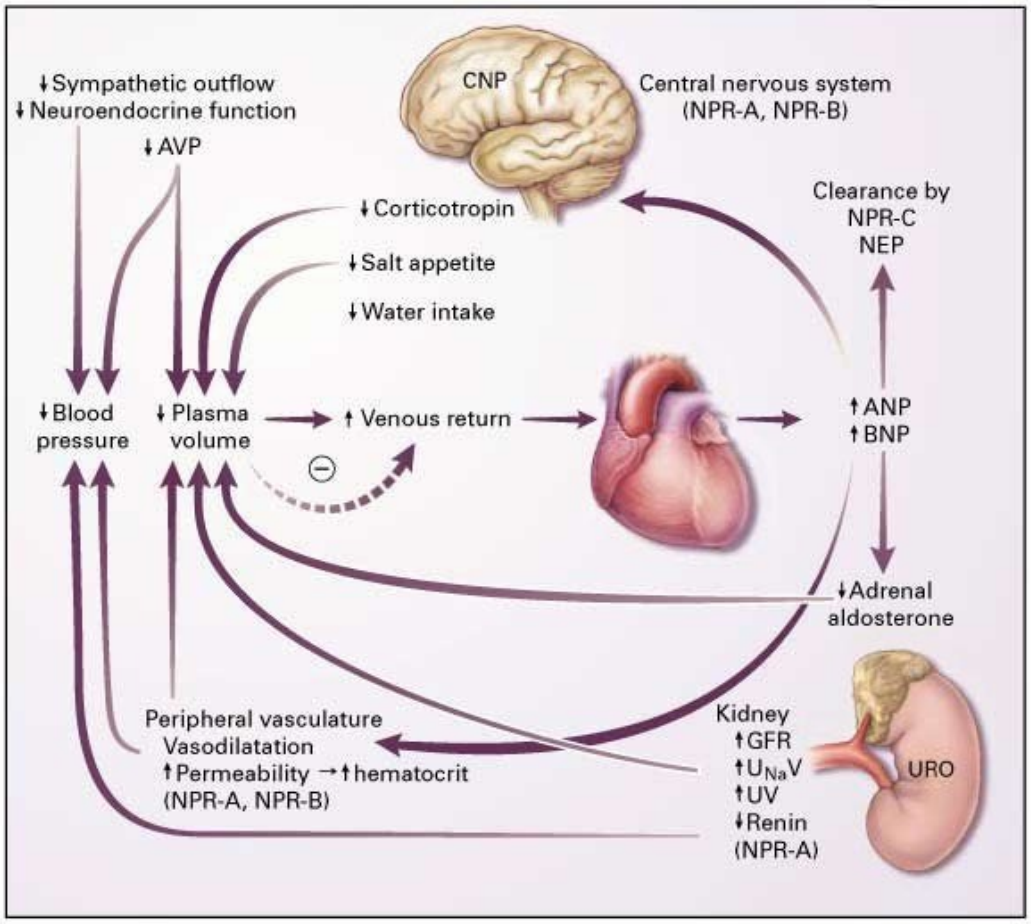


FIGURE 1.5. PHYSIOLOGIC ACTIONS OF CIRCULATING NATRIURETIC PEPTIDES, INCLUDING BNP [106]

1.3.3 BNP EPIDEMIOLOGY AND ENVIRONMENTAL EPIDEMIOLOGY

What is well understood at this point is that BNP is an important diagnostic and prognostic marker in heart failure patients [28, 119]. Beyond diagnosis, BNP is increasingly linked to survival in epidemiologic studies of patients with and without cardiovascular disease [32, 120-122]. This association can exist even in asymptomatic patients, indicating that BNP can be a subclinical marker of intracardiac pressure and of the complex cascade of biological processes that follow [31, 123]. Although epidemiologic associations between BNP levels and survival are fairly well established, there are very few epidemiologic studies that evaluate BNP with respect to

environmental factors [124-126]. Of these studies, associations between environmental factors (e.g., air pollution) and BNP are inconsistent, and are likely due to the fact that these studies had very small sample sizes (i.e., n = 28 – 45). Epidemiologic evaluation of environmental factors in relation to BNP is thus a very fruitful area of research, because it is a molecule with important prognostic significance in heart failure; it might shed light on mechanistic pathways; it is commonly and regularly measured in subjects with heart failure, and thus may minimize misclassification issues associated with the identification of heart failure exacerbation by hospitalization alone; and few studies of environmental exposures in relation to BNP have been completed to date.

1.4 Environmental Epidemiology of Heart Failure

1.4.1 Environmental exposure pathways relevant to heart failure hospitalizations

Exposure to air pollution, and to particulate matter specifically, has been associated with adverse cardiovascular outcomes, as reviewed in Rajagopalan et al. [127, 128]. The associations between air pollution and cardiovascular disease exist in the short-term (i.e., exacerbations and hospitalizations) [129, 130] as well as in the long term (i.e., contributing to the progression of cardiac disease and mortality) [131-133]. Mechanisms underlying these associations include: systemic inflammation [134, 135], endothelial dysfunction [136, 137], blood coagulation [138, 139], and consequential cardiac remodeling [140, 141]. Observational studies that elucidate these mechanisms, however, are limited by the inability to fully characterize the effects of comorbidities that could mediate or modify the associations between environmental factors and cardiovascular outcomes.

1.4.2 Environmental Epidemiology of Heart Failure hospitalizations

Heart failure is an excellent health outcome for study in relation to UNGD because prior studies suggest that heart failure can be exacerbated by toxicants in air pollution [142] and stress [143], two primary sets of UNGD impacts. There is consistent

epidemiologic evidence that supports links between air pollution and heart failure; a recent review by Shah et al., found increases in heart failure hospital admissions and mortality on the day of and the day following increases in $PM_{2.5}$, SO_x, NO_x, and CO [142]. Studies of the short-term effects of air pollution find similar associations for both cardiovascular and respiratory outcomes [144, 145], and both cardiovascular and respiratory exacerbations are influenced by stress [146-148]. Because no studies have yet evaluated associations between UNGD activity and any health outcome in older adults, and because persons living with heart failure might be susceptible to the air pollution and community impacts associated with UNGD, we evaluated associations between UNGD activity and heart failure outcomes. In the next sections, I describe the environmental epidemiology and rationale for also considering additional environmental and community variables in this research.

1.4.3 Greenness

A growing body of epidemiological evidence reports that greenness is associated with beneficial health effects [50, 51, 149, 150]. These studies find that the normalized difference vegetative index (NDVI), a common measure of greenness, is associated with decreased risk of adverse birth outcomes (e.g., preterm birth, birth weight, small for gestational age [SGA]), diabetes, hypertension and hyperlipidemia, cardiovascular disease, mental health conditions, and mortality [50, 51, 57, 149, 151, 152]. Proposed mechanisms for NDVI-health associations include: reducing air and noise pollution, providing opportunities for physical activity, providing opportunities for social interaction, and through direct improvements in mental health conditions such as stress and depression [149, 151], particularly from the perspective of providing mental restoration and stress reduction [153, 154]. Other studies that measured salivary cortisol, heart rate, and cognitive function before and after individuals' exposure to natural environments

with higher levels of greenness documented acute health benefits (i.e., reduced salivary cortisol and heart rate, and improved cognitive function) in response to spending as little as 30 minutes in these environments [153, 155]. Similarly, a recent longitudinal study of NDVI and mortality in the Nurses' Health Study cohort (women only) found that NDVI, calculated at a 250m buffer around a participants' residence, was independently associated with reduced non-accidental mortality and this association was, in part, mediated by measures of physical activity, depression, $PM_{2.5}$, and social engagement [51], and the authors estimated that physician-diagnosed depression or anti-depressive medication use explained 30.6% (95% CI: 15.5%, 51.4%) of the association between NDVI and non-accidental mortality [51]. Another study utilized the natural experiment of the emerald ash borer, an invasive pest which killed over 100 million ash trees between the years 2002-2010 in the Midwestern US, and its associations with increased risk of cardiovascular disease in women [156]. This study documented increased hazard (hazard ratio: 1.2, 95% CI: 1.20-1.31) for cardiovascular disease development, comparing women living in counties that were infested with the emerald ash borer to those living in counties that were not infested [156]. NDVI was thus included in this research because we suspected that it could be associated with heart failure hospitalization and levels of BNP in heart failure subjects.

1.4.4 Community socioeconomic deprivation

Epidemiologic studies demonstrate that community socioeconomic deprivation (CSD) is independently associated with cardiovascular diseases, diabetes, body mass index (BMI), cognitive function, and overall morbidity and mortality [157-161]. Understanding the mechanisms responsible for these associations is difficult because CSD is likely confounded by a number of individual- (e.g., race/ethnicity, socioeconomic status [SES], economic disadvantage, fear of crime) and community-level exposures

(i.e., racial segregation, lack of community resources) that can work through a variety of biological pathways to cause these adverse health outcomes [162, 163]. These stressors can act to impact physical activity and social engagement, exacerbating the effects of CSD on health [164-166], either through epidemiologic mediation or effect modification. In addition, many of these factors are associated with mental health comorbidities (e.g., depression [167, 168]) and could therefore be associated with heart failure etiology and susceptibility to heart failure hospitalization. This is consistent with studies that, although unable to disentangle mechanisms, provide growing evidence that deprivation measures are independently associated with heart failure hospitalization and mortality [169, 170]. For these reasons, we evaluated CSD in relation to all of the heart failure outcomes considered in this research, and in relation to UNGD activity and NDVI to assess the extent to which CSD could moderate or mediate associations between UNGD activity, NDVI, and heart failure outcomes.

1.4.5 Proximity to roadways

A substantial body of epidemiologic literature has examined residential proximity to major and minor roadways in relation to adverse health outcomes, including hypertension [171], low birth weight [172], type 2 diabetes [173], and markers of inflammation [174]. Often, proximity to roadways is used as a proxy for air pollution and noise exposure due to traffic, although wind and atmospheric conditions can lead to differential dispersion of noise and air pollution from these sources [175]. Despite the uncertainty in etiological agent, residential proximity to roadways can be a good indicator of both air pollution and noise for use in epidemiologic studies of cardiovascular outcomes [176, 177]. Although this was not the main environmental exposure considered in this research, we evaluated the associations between proximity to major and minor roadways in analyses of heart failure hospitalization and with BNP levels

because it could confound associations between UNGD and heart failure outcomes; community deprivation and heart failure outcomes; and NDVI and heart failure outcomes.

1.5 Public Health Relevance

This study is the first epidemiologic study to examine associations between UNGD activity and heart failure hospitalization – the first evaluation of any health outcome in relation to UNGD activity that primarily affects older adults. Given growing evidence that UNGD is associated with several health impacts in the Marcellus shale (e.g., asthma exacerbation [15], migraine headaches and fatigue [17], birth outcomes [9-11, 13, 64], reports of stress [63], depressive symptoms, and disordered sleep [18]), we suspected that older adults, and particularly those with heart failure, would likely be most susceptible to adverse environmental exposures and contextual effects of UNGD activity. Importantly, this study is the first to evaluate associations between UNGD activity and any biological marker in a large, representative population of subjects with heart failure. Adding to the epidemiologic literature regarding UNGD and health impacts, the evaluation of BNP in relation to UNGD activity is the first study to evaluate environmental factors of any kind in relation to BNP levels in a large population.

To the best of our knowledge, this study is first epidemiologic study to examine effect modification of environmental associations with heart failure hospitalization and with BNP levels by HF*p*EF and HF*r*EF phenotype status of the associations between UNGD and hospitalization and the association between UNGD activity and BNP. Being able to systematically distinguish heart failure phenotypes has been a limitation in previous large scale epidemiology studies of heart failure, but advances in data science have made extracting heart failure phenotype information from EHRs a novel possibility [178, 179]. This study evaluated whether heart failure phenotype modified the

associations between UNGD activity and hospitalization and UNGD activity and BNP levels because HF*p*EF and HF*r*EF are two distinct phenotypes of heart failure with differing pathophysiology. We suspected that understanding effect modification of these two associations by phenotypes would provide insights to biologic mechanisms underlying these associations.

We hypothesized that environmental associations would be stronger among HF*p*EF subjects than among HFrEF subjects because heart failure, and HF*p*EF in particular, is increasing in prevalence and is more commonly associated with non-cardiac comorbidities that are related to the built and natural environment than HF*r*EF. This possibility has been posited by several studies of differential responses to stressors (chemical and non-chemical) in heart failure exacerbation (e.g., mortality, hospitalization) by phenotype [180, 181]. Greater susceptibility to environmental factors related to exacerbation by phenotype would provide insights beneficial to understanding the etiology and treatment of heart failure and could help explain recent trends in increasing HF*p*EF prevalence relative to HF*r*EF.

1.6 SPECIFIC AIMS

Given the aforementioned considerations, the specific aims of this dissertation research were to:

- SA1 (**Chapter 3)**: Evaluate associations of four phases of UNGD activity with heart failure hospitalizations in a case-control analysis (i.e., comparing heart failure patients with and without heart failure hospitalizations) using multilevel logistic regression.
- SA2 (**Chapter 4**): Evaluate associations of four phases of UNGD activity with BNP levels in blood among subjects with heart failure, using generalized estimating equations (GEE).

- SA3 (**Chapter 5**): Evaluate whether heart failure phenotype (HF*p*EF and HF*r*EF) status modified associations of UNGD activity with heart failure hospitalization and BNP levels in separate analyses:
    - SA3a (**Chapter 5a**): Evaluate effect modification by phenotype status of the association between UNGD activity and heart failure hospitalization.
    - SA3b (**Chapter 5b**): Evaluate effect modification by phenotype status of the association between UNGD activity and BNP levels.

1.7 REFERENCES

1. Werner, A.K., et al., *Environmental health impacts of unconventional natural gas development: A review of the current strength of evidence.* Science of The Total Environment, 2015. **505**: p. 1127-1141.
2. Vengosh, A., et al., *A Critical Review of the Risks to Water Resources from Unconventional Shale Gas Development and Hydraulic Fracturing in the United States.* Environmental Science & Technology, 2014. **48**(15): p. 8334-8348.
3. Hays, J. and S.B.C. Shonkoff, *Toward an Understanding of the Environmental and Public Health Impacts of Unconventional Natural Gas Development: A Categorical Assessment of the Peer-Reviewed Scientific Literature, 2009-2015.* PLoS ONE, 2016. **11**(4): p. e0154164.
4. Boyle, M.D., et al., *A pilot study to assess residential noise exposure near natural gas compressor stations.* PloS one, 2017. **12**(4): p. e0174310-e0174310.
5. Hoelzer, K., et al., *Indications of Transformation Products from Hydraulic Fracturing Additives in Shale-Gas Wastewater.* Environmental Science & Technology, 2016.
6. Graham, J., et al., *Increased traffic accident rates associated with shale gas drilling in Pennsylvania.* Accident Analysis & Prevention, 2015. **74**: p. 203-209.
7. Gopalakrishnan, S. and H.A. Klaiber, *Is the Shale Energy Boom a Bust for Nearby Residents? Evidence from Housing Values in Pennsylvania.* American Journal of Agricultural Economics, 2014. **96**(1): p. 43-66.
8. Komarek, T.M., *Crime and natural resource booms: evidence from unconventional natural gas production.* The Annals of Regional Science, 2018. **61**(1): p. 113-137.
9. Stacy, S., et al., *Perinatal outcomes and unconventional natural gas operations in Southwest Pennsylvania.* PLoS One, 2015. **10**(6): p. e0126425.
10. Currie, J., M. Greenstone, and K. Meckel, *Hydraulic fracturing and infant health: New evidence from Pennsylvania.* Science Advances, 2017. **3**(12): p. e1603021.
11. Whitworth, K.W., A.K. Marshall, and E. Symanski, *Maternal residential proximity to unconventional gas development and perinatal outcomes among a diverse urban population in Texas.* PloS one, 2017. **12**(7): p. e0180966-e0180966.
12. Walker Whitworth, K., A. Kaye Marshall, and E. Symanski, *Drilling and Production Activity Related to Unconventional Gas Development and Severity of Preterm Birth.* 2018(1552-9924 (Electronic)).
13. Casey, J.A., et al., *Unconventional Natural Gas Development and Birth Outcomes in Pennsylvania, USA*. Epidemiology, 2015.
14. Janitz, A.E., et al., *The association between natural gas well activity and specific congenital anomalies in Oklahoma, 1997–2009.* Environment International, 2019. **122**: p. 381-388.
15. Rasmussen, S.G., et al., *Association Between Unconventional Natural Gas Development in the Marcellus Shale and Asthma Exacerbations. .* JAMA Internal Medicine, 2016(2168-6114).
16. Willis, M.D., et al., *Unconventional natural gas development and pediatric asthma hospitalizations in Pennsylvania.* (1096-0953 (Electronic)).
17. Tustin, A.W., et al., *Associations between Unconventional Natural Gas Development and Nasal and Sinus, Migraine Headache, and Fatigue Symptoms in Pennsylvania.* Environ Health Perspect, 2016.
18. Casey, J.A., et al., *Associations of unconventional natural gas development with depression symptoms and disordered sleep in Pennsylvania.* Scientific Reports, 2018. **8**(1): p. 11375.

19. Go Alan, S., et al., *Heart Disease and Stroke Statistics—2013 Update.* Circulation, 2013. **127**(1): p. e6-e245.
20. Jackson Sandra, L., et al., *National Burden of Heart Failure Events in the United States, 2006 to 2014.* Circulation: Heart Failure, 2018. **11**(12): p. e004873.
21. Heidenreich, P.A., et al., *Forecasting the Future of Cardiovascular Disease in the United States.* Circulation, 2011. **123**(8): p. 933.
22. McKenzie, L.M., et al., *Human health risk assessment of air emissions from development of unconventional natural gas resources.* Sci Total Environ, 2012. **424**: p. 79-87.
23. Maryland Department of Environment and Maryland Department of Natural Resources, *Assessment of Risks from Unconventional Gas Well Development in the Marcellus Shale of Western Maryland*. 2015.
24. Swarthout, R.F., et al., *Impact of Marcellus Shale Natural Gas Development in Southwest Pennsylvania on Volatile Organic Compound Emissions and Regional Air Quality.* Environmental Science & Technology, 2015. **49**(5): p. 3175-3184.
25. Johnson, D., et al., *Estimated Emissions from the Prime-Movers of Unconventional Natural Gas Well Development Using Recently Collected In-Use Data in the United States.* Environmental Science & Technology, 2018.
26. Shah, A.S.V., et al., *Global association of air pollution and heart failure: a systematic review and meta-analysis.* The Lancet. **382**(9897): p. 1039-1048.
27. Adgate, J.L., B.D. Goldstein, and L.M. McKenzie, *Potential Public Health Hazards, Exposures and Health Effects from Unconventional Natural Gas Development.* Environmental Science & Technology, 2014. **48**(15): p. 8307-8320.
28. Magnussen, C. and S. Blankenberg, *Biomarkers for heart failure: small molecules with high clinical relevance.* Journal of Internal Medicine, 2018. **283**(6): p. 530-543.
29. Wolsk, E., et al., *Role of B-Type Natriuretic Peptide and N-Terminal Prohormone BNP as Predictors of Cardiovascular Morbidity and Mortality in Patients With a Recent Coronary Event and Type 2 Diabetes Mellitus.* Journal of the American Heart Association. **6**(6): p. e004743.
30. van Veldhuisen, D.J., et al., *B-Type Natriuretic Peptide and Prognosis in Heart Failure Patients With Preserved and Reduced Ejection Fraction.* Journal of the American College of Cardiology, 2013. **61**(14): p. 1498-1506.
31. York, M.K., et al., *B-Type Natriuretic Peptide Levels and Mortality in Patients With and Without Heart Failure.* Journal of the American College of Cardiology, 2018. **71**(19): p. 2079-2088.
32. Zhang, C., et al., *Brain Natriuretic Peptide as the long-term cause of mortality in patients with cardiovascular disease: a retrospective cohort study.* International journal of clinical and experimental medicine, 2015. **8**(9): p. 16364-16368.
33. US EIA, *Annual Energy Outlook 2016* U.E.I. Association, Editor. 2016, US EIA.
34. The Nature Conservancy, *Pennsylvania Energy Impacts Assessment, Report 1: Marcellus shale natural gas and wind.* 2010, The Nature Conservancy -- Pennsylvania Chapter.
35. Board on Chemical Sciences and Technology; Division on Earth and Life Studies; National Academies of Sciences, E., and Medicine;, *Chemistry and Engineering of Shale Gas and Tight Oil Resources Development: Workshop in Brief*. 2015, National Academies of Sciences, Engineering, and Medicine;. p. 16.
36. King, G.E., *Hydraulic Fracturing 101: What Every Representative, Environmentalist, Regulator, Reporter, Investor, University Researcher, Neighbor*

and Engineer Should Know About Estimating Frac Risk and Improving Frac Performance in Unconventional Gas and Oil Wells. 2012: Woodlands, TX.
37. Brown Valeria, J., *Radionuclides in Fracking Wastewater: Managing a Toxic Blend.* Environmental Health Perspectives, 2014. **122**(2): p. A50-A55.
38. Field, R.A., J. Soltis, and S. Murphy, *Air quality concerns of unconventional oil and natural gas production.* Environmental Science: Processes & Impacts, 2014. **16**(5): p. 954-969.
39. Litovitz, A., et al., *Estimation of regional air-quality damages from Marcellus Shale natural gas extraction in Pennsylvania.* Environmental Research Letters, 2013. **8**(1): p. 014017.
40. Roy, A.A., P.J. Adams, and A.L. Robinson, *Air pollutant emissions from the development, production, and processing of Marcellus Shale natural gas.* Journal of the Air & Waste Management Association, 2014. **64**(1): p. 19-37.
41. Brown, D.R., C. Lewis, and B.I. Weinberger, *Human exposure to unconventional natural gas development: A public health demonstration of periodic high exposure to chemical mixtures in ambient air.* Journal of Environmental Science and Health, Part A, 2015. **50**(5): p. 460-472.
42. Elliott, E.G., et al., *A systematic evaluation of chemicals in hydraulic-fracturing fluids and wastewater for reproductive and developmental toxicity.* Journal of Exposure Science and Environmental Epidemiology, 2016.
43. Council, N.R.D., *Fracking 101.* 2019.
44. Paulik, L.B., et al., *Emissions of Polycyclic Aromatic Hydrocarbons from Natural Gas Extraction into Air.* Environmental science & technology, 2016. **50**(14): p. 7921-7929.
45. Osborn, S.G., et al., *Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing.* Proceedings of the National Academy of Sciences, 2011. **108**(20): p. 8172-8176.
46. Osborn, S.G., et al., *Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(20): p. 8172-8176.
47. Loveless, S.E., et al., *Characterising the vertical separation of shale-gas source rocks and aquifers across England and Wales (UK).* Hydrogeology journal, 2018. **26**(6): p. 1975-1987.
48. Moore, C.W., et al., *Air Impacts of Increased Natural Gas Acquisition, Processing, and Use: A Critical Review.* Environmental Science & Technology, 2014. **48**(15): p. 8349-8359.
49. Morelli, X., et al., *Short-term associations between traffic-related noise, particle number and traffic flow in three European cities.* Atmospheric Environment, 2015. **103**: p. 25-33.
50. James, P., et al., *A Review of the Health Benefits of Greenness.* Current Epidemiology Reports, 2015. **2**(2): p. 131-142.
51. James, P., et al., *Exposure to Greenness and Mortality in a Nationwide Prospective Cohort Study of Women.* Environ Health Perspect, 2016. **124**(9): p. 1344-52.
52. Wang, K., et al., *Relationship of Neighborhood Greenness to Heart Disease in 249 405 US Medicare Beneficiaries.* Journal of the American Heart Association, 2019. **8**(6): p. e010258-e010258.
53. Field, R.A., S. Soltis J Fau - Murphy, and S. Murphy, *Air quality concerns of unconventional oil and natural gas production.* 2014(2050-7895 (Electronic)).

54. Perry, S.L., *Development, Land Use, and Collective Trauma: The Marcellus Shale Gas Boom in Rural Pennsylvania.* Culture, Agriculture, Food and Environment, 2012. **34**(1): p. 81-92.
55. Muehlenbachs, L., E. Spiller, and C. Timmins, *The Housing Market Impacts of Shale Gas Development.* American Economic Review, 2015. **105**(12): p. 3633-59.
56. EIA;, U., *Shale plays in the Lower 48 states.* 2015.
57. Casey, J.A., et al., *Greenness and Birth Outcomes in a Range of Pennsylvania Communities.* International Journal of Environmental Research and Public Health, 2016. **13**(3).
58. US Department of Energy. *How Is Shale Gas Produced?* 2013 [cited 2015 December]; Available from: http://energy.gov/sites/prod/files/2013/04/f0/how_is_shale_gas_produced.pdf.
59. Powers, M., et al., *Popular epidemiology and "fracking": citizens' concerns regarding the economic, environmental, health and social impacts of unconventional natural gas drilling operations.* (1573-3610 (Electronic)).
60. Union of Concerned Scientists, *Toward an Evidence-Based Fracking Debate: Science, Democracy, and Community Right to Know in Unconventional Oil and Gas Development*. 2013.
61. Pennsylvania Department of Environmental Protection, *PA DEP Oil & Gas Reporting Website* 2016.
62. Casey, J., A., et al., *Predictors of Indoor Radon Concentrations in Pennsylvania, 1989–2013.* Environmental Health Perspectives, 2015. **123**(11): p. 1130-1137.
63. Rabinowitz, P.M., et al., *Proximity to natural gas wells and reported health status: results of a household survey in Washington County, Pennsylvania.* Environ Health Perspect, 2015. **123**(1): p. 21-6.
64. McKenzie, L.M., et al., *Birth outcomes and maternal residential proximity to natural gas development in rural Colorado.* Environmental health perspectives, 2014. **122**(4): p. 412-417.
65. Walker, W.K., K.A. Marshall, and E. Symanski, *Drilling and Production Activity Related to Unconventional Gas Development and Severity of Preterm Birth.* Environ Health Perspect. , 2018. **126**(3): p. 037006.
66. McKenzie, L.M., et al., *Relationships between indicators of cardiovascular disease and intensity of oil and natural gas activity in Northeastern Colorado.* Environmental Research, 2019. **170**: p. 56-64.
67. Bujak, M. and N.G. Frangogiannis, *The role of IL-1 in the pathogenesis of heart disease.* (1661-4917 (Electronic)).
68. Szekely, Y. and Y. Arbel, *A Review of Interleukin-1 in Heart Disease: Where Do We Stand Today?* Cardiology and therapy, 2018. **7**(1): p. 25-44.
69. Yu, Y., et al., *TNF-α receptor 1 knockdown in the subfornical organ ameliorates sympathetic excitation and cardiac hemodynamics in heart failure rats.* American journal of physiology. Heart and circulatory physiology, 2017. **313**(4): p. H744-H756.
70. Zhang, H., et al., *Role of TNF-alpha in vascular dysfunction.* (1470-8736 (Electronic)).
71. Vasan Ramachandran, S., et al., *Inflammatory Markers and Risk of Heart Failure in Elderly Subjects Without Prior Myocardial Infarction.* Circulation, 2003. **107**(11): p. 1486-1491.
72. Kalogeropoulos, A., et al., *Inflammatory Markers and Incident Heart Failure Risk in Older Adults.* Journal of the American College of Cardiology, 2010. **55**(19): p. 2129.

73. CDC;. *Heart Failure Fact Sheet*. 2013 [cited 2015 October]; Available from: http://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_failure.htm.
74. Heidenreich, P.A., et al., *Forecasting the Impact of Heart Failure in the United States: A Policy Statement From the American Heart Association.* Circulation: Heart Failure, 2013.
75. Mozaffarian, D., et al., *Heart Disease and Stroke Statistics—2016 Update: A Report From the American Heart Association.* Circulation, 2016. **133**(4): p. e38-e360.
76. Cowie, M.R., et al., *Survival of patients with a new diagnosis of heart failure: a population based study.* (1468-201X (Electronic)).
77. Roger, V.L., *Epidemiology of heart failure.* Circulation research, 2013. **113**(6): p. 646-659.
78. Heo, S., et al., *Modifiable correlates of physical symptoms and health-related quality of life in patients with heart failure: a cross-sectional study.* (1873-491X (Electronic)).
79. Lucas, R., et al., *The effect of heart failure nurse consultations on heart failure patients' illness beliefs, mood and quality of life over a six-month period.* 2015(1365-2702 (Electronic)).
80. Nieminen, M.S., et al., *The patient perspective: Quality of life in advanced heart failure with frequent hospitalisations.* International Journal of Cardiology, 2015. **191**: p. 256-264.
81. Lee, D.S., et al., *Relation of Disease Pathogenesis and Risk Factors to Heart Failure With Preserved or Reduced Ejection Fraction.* Circulation, 2009. **119**(24): p. 3070-3077.
82. Borlaug, B.A., *Heart failure with preserved and reduced ejection fraction: different risk profiles for different diseases.* European Heart Journal, 2013.
83. Udelson, J.E., *Heart Failure With Preserved Ejection Fraction.* Circulation, 2011. **124**(21): p. e540-e543.
84. Dharmarajan, K., et al., *Trajectories of risk after hospitalization for heart failure, acute myocardial infarction, or pneumonia: retrospective cohort study.* BMJ, 2015. **350**.
85. Brouwers, F.P., et al., *Incidence and epidemiology of new onset heart failure with preserved vs. reduced ejection fraction in a community-based cohort: 11-year follow-up of PREVEND.* European Heart Journal, 2013. **34**(19): p. 1424-1431.
86. Salamanca-Bautista, P., et al., *Precipitating factors of heart failure admission: Differences related to age and left ventricular ejection fraction.* (1874-1754 (Electronic)).
87. Dominguez-Rodriguez, A., et al., *Air pollution and heart failure: Relationship with the ejection fraction.* World Journal of Cardiology, 2013. **5**(3): p. 49-53.
88. Owan, T.E., et al., *Trends in Prevalence and Outcome of Heart Failure with Preserved Ejection Fraction.* New England Journal of Medicine, 2006. **355**(3): p. 251-259.
89. Bui, A.L., T.B. Horwich, and G.C. Fonarow, *Epidemiology and risk profile of heart failure.* Nature reviews. Cardiology, 2011. **8**(1): p. 30-41.
90. Paulus, W.J. and C. Tschöpe, *A Novel Paradigm for Heart Failure With Preserved Ejection Fraction: Comorbidities Drive Myocardial Dysfunction and Remodeling Through Coronary Microvascular Endothelial Inflammation.* Journal of the American College of Cardiology, 2013. **62**(4): p. 263-271.
91. Shah, S.J., D.H. Katz, and R.C. Deo, *Phenotypic Spectrum of Heart Failure with Preserved Ejection Fraction.* Heart failure clinics, 2014. **10**(3): p. 407-418.

92. De Keulenaer Gilles, W. and L. Brutsaert Dirk, *Systolic and Diastolic Heart Failure Are Overlapping Phenotypes Within the Heart Failure Spectrum.* Circulation, 2011. **123**(18): p. 1996-2005.
93. Poornima Indu, G., P. Parikh, and P. Shannon Richard, *Diabetic Cardiomyopathy.* Circulation Research, 2006. **98**(5): p. 596-605.
94. Wang, J. and T. Guo, *Metabolic remodeling in chronic heart failure.* (1862-1783 (Electronic)).
95. Abel, E.D., G. Litwin Se Fau - Sweeney, and G. Sweeney, *Cardiac remodeling in obesity.* (0031-9333 (Print)).
96. Mudd, J.O. and D.A. Kass, *Tackling heart failure in the twenty-first century.* (1476-4687 (Electronic)).
97. De Keulenaer Gilles, W. and L. Brutsaert Dirk, *The Heart Failure Spectrum.* Circulation, 2009. **119**(24): p. 3044-3046.
98. Potter, L.R., et al., *Natriuretic peptides: their structures, receptors, physiologic functions and therapeutic applications.* Handbook of experimental pharmacology, 2009(191): p. 341-366.
99. de Lemos, J.A., M.H. McGuire Dk Fau - Drazner, and M.H. Drazner, *B-type natriuretic peptide in cardiovascular disease.* (1474-547X (Electronic)).
100. Maries, L. and I. Manitiu, *Diagnostic and prognostic values of B-type natriuretic peptides (BNP) and N-terminal fragment brain natriuretic peptides (NT-pro-BNP).* (1680-0745 (Electronic)).
101. Felker, G.M., et al., *Effect of Natriuretic Peptide–Guided Therapy on Hospitalization or Cardiovascular Mortality in High-Risk Patients With Heart Failure and Reduced Ejection Fraction: A Randomized Clinical TrialNatriuretic Peptide–Guided Therapy for Heart FailureNatriuretic Peptide–Guided Therapy for Heart Failure.* JAMA, 2017. **318**(8): p. 713-720.
102. Ibrahim, N. and J.L. Januzzi, *The potential role of natriuretic peptides and other biomarkers in heart failure diagnosis, prognosis and management.* (1744-8344 (Electronic)).
103. Fonseca, C., et al., *Comparative value of BNP and NT-proBNP in diagnosis of heart failure.* (0870-2551 (Print)).
104. Laboratories, G.M. *Geisinger Medical Laboratories Test Catalog*. 2019; Available from: https://www.geisingermedicallabs.com/catalog/details.cfm?tid=1719.
105. Weber, M., C. Mitrovic V Fau - Hamm, and C. Hamm, *B-type natriuretic peptide and N-terminal pro-B-type natriuretic peptide - Diagnostic role in stable coronary artery disease.* (1205-6626 (Print)).
106. Levin, E.R., D.G. Gardner, and W.K. Samson, *Natriuretic Peptides.* New England Journal of Medicine, 1998. **339**(5): p. 321-328.
107. Potter, L.R., *Natriuretic peptide metabolism, clearance and degradation.* The FEBS journal, 2011. **278**(11): p. 1808-1817.
108. Schulz, S., et al., *The primary structure of a plasma membrane guanylate cyclase demonstrates diversity within this new receptor family.* (0092-8674 (Print)).
109. Ogawa, Y., et al., *Molecular cloning of the complementary DNA and gene that encode mouse brain natriuretic peptide and generation of transgenic mice that overexpress the brain natriuretic peptide gene.* (0021-9738 (Print)).
110. Clerico, A., *Pathophysiological and clinical relevance of circulating levels of cardiac natriuretic hormones: are they merely markers of cardiac disease?* (1434-6621 (Print)).

111. Kjær, A. and B. Hesse, *Heart failure and neuroendocrine activation: diagnostic, prognostic and therapeutic perspectives.* Clinical Physiology, 2001. **21**(6): p. 661-672.
112. Popovic, D., et al., *The interface of hypothalamic-pituitary-adrenocortical axis and circulating brain natriuretic peptide in prediction of cardiopulmonary performance during physical stress.* (1873-5169 (Electronic)).
113. Wiedemann, K., M. Jahn H Fau - Kellner, and M. Kellner, *Effects of natriuretic peptides upon hypothalamo-pituitary-adrenocortical system activity and anxiety behaviour.* (0947-7349 (Print)).
114. Amir, O., et al., *The response of circulating brain natriuretic peptide to academic stress in college students.* (1607-8888 (Electronic)).
115. Wirtz, P.H., et al., *Increases in B-type natriuretic peptide after acute mental stress in heart failure patients are associated with alcohol consumption.* (1938-4114 (Electronic)).
116. Laederach-Hofmann, K., et al., *Comprehensive rehabilitation in chronic heart failure--better psycho-emotional status related to quality of life, brain natriuretic peptide concentrations, and clinical severity of disease.* (1488-2353 (Electronic)).
117. Sudoh, T., et al., *A new natriuretic peptide in porcine brain.* (0028-0836 (Print)).
118. Mair, J., *Biochemistry of B-type natriuretic peptide--where are we now?* (1434-6621 (Print)).
119. Yancy, C.W., et al., *2017 ACC/AHA/HFSA Focused Update of the 2013 ACCF/AHA Guideline for the Management of Heart Failure.* Journal of the American College of Cardiology, 2017. **70**(6): p. 776.
120. Santaguida, P.L., et al., *BNP and NT-proBNP as prognostic markers in persons with acute decompensated heart failure: a systematic review.* (1573-7322 (Electronic)).
121. York, M.K., et al., *B-Type Natriuretic Peptide Levels and Mortality in Patients With and Without Heart Failure.* (1558-3597 (Electronic)).
122. Khanam, S.S., et al., *Prognostic value of short-term follow-up BNP in hospitalized patients with heart failure.* (1471-2261 (Electronic)).
123. Vodovar, N. and D. Logeart, *Similar BNP and Mortality Association in Patients With and Without Heart Failure.* Journal of the American College of Cardiology, 2018. **71**(19): p. 2089.
124. Wellenius, G.A., et al., *Effects of ambient air pollution on functional status in patients with chronic congestive heart failure: a repeated-measures study.* (1476-069X (Electronic)).
125. Vieira, J.L., et al., *Respiratory Filter Reduces the Cardiovascular Effects Associated With Diesel Exhaust Exposure: A Randomized, Prospective, Double-Blind, Controlled Study of Heart Failure: The FILTER-HF Trial.* (2213-1787 (Electronic)).
126. Emiroglu, Y., et al., *BNP levels in patients with long-term exposure to biomass fuel and its relation to right ventricular function.* (1522-9629 (Electronic)).
127. Mills, N.L., et al., *Adverse cardiovascular effects of air pollution.* Nat Clin Pract Cardiovasc Med, 2009. **6**(1): p. 36-44.
128. Rajagopalan, S., S.G. Al-Kindi, and R.D. Brook, *Air Pollution and Cardiovascular Disease: JACC State-of-the-Art Review.* (1558-3597 (Electronic)).
129. Rodopoulou, S., et al., *Air pollution and cardiovascular and respiratory emergency visits in Central Arkansas: A time-series analysis.* (1879-1026 (Electronic)).

130. Wellenius, G.A., et al., *Particulate Air Pollution and the Rate of Hospitalization for Congestive Heart Failure among Medicare Beneficiaries in Pittsburgh, Pennsylvania.* American Journal of Epidemiology, 2005. **161**(11): p. 1030-1036.
131. Pope, C.A., et al., *Relationships Between Fine Particulate Air Pollution, Cardiometabolic Disorders, and Cardiovascular Mortality.* Circulation Research, 2015. **116**(1): p. 108-115.
132. Cai, Y., et al., *Long-term exposure to road traffic noise, ambient air pollution, and cardiovascular risk factors in the HUNT and lifelines cohorts.* (1522-9645 (Electronic)).
133. Kim, H., et al., *Cardiovascular Effects of Long-Term Exposure to Air Pollution: A Population-Based Study With 900 845 Person-Years of Follow-up. LID - e007170 [pii] LID - 10.1161/JAHA.117.007170 [doi].* (2047-9980 (Electronic)).
134. Li, W., et al., *Short-Term Exposure to Ambient Air Pollution and Biomarkers of Systemic Inflammation.* Arteriosclerosis, Thrombosis, and Vascular Biology, 2017. **37**(9): p. 1793-1800.
135. Siponen, T., et al., *Source-specific fine particulate air pollution and systemic inflammation in ischaemic heart disease patients.* (1470-7926 (Electronic)).
136. Bourdrel, T., et al., *Cardiovascular effects of air pollution.* (1875-2128 (Electronic)).
137. Hwang, S.J., et al., *Endothelial dysfunction in heart failure rats exposed to real urban air pollution.* (1874-1754 (Electronic)).
138. Viehmann, A., et al., *Long-term residential exposure to urban air pollution, and repeated measures of systemic blood markers of inflammation and coagulation.* (1470-7926 (Electronic)).
139. Dominguez-Rodriguez, A., et al., *Short-term effects of air pollution, markers of endothelial activation, and coagulation to predict major adverse cardiovascular events in patients with acute coronary syndrome: insights from AIRACOS study.* (1366-5804 (Electronic)).
140. Liu, Y., et al., *Air pollution and adverse cardiac remodeling: clinical effects and basic mechanisms.* (1664-042X (Print)).
141. de Oliveira-Fonoff, A.M., et al., *The role of air pollution in myocardial remodeling.* (1932-6203 (Electronic)).
142. Shah, A.S., et al., *Short term exposure to air pollution and stroke: systematic review and meta-analysis.* (1756-1833 (Electronic)).
143. Ogilvie, R.P., et al., *Psychosocial Factors and Risk of Incident Heart Failure: The Multi-Ethnic Study of Atherosclerosis.* Circ Heart Fail, 2016. **9**(1): p. e002243.
144. Fryzek, J., et al., *Childhood cancer incidence in Pennsylvania counties in relation to living in counties with hydraulic fracturing sites.* (1536-5948 (Electronic)).
145. Mann, J.K., et al., *Short-Term Effects of Air Pollution on Wheeze in Asthmatic Children in Fresno, California.* Environmental Health Perspectives, 2010. **118**(10): p. 1497-1502.
146. Chen, E., et al., *Chronic Traffic-Related Air Pollution and Stress Interact to Predict Biologic and Clinical Outcomes in Asthma.* 2008.
147. McDowell, K.M., et al., *Medical and Social Determinants of Health Associated with Intensive Care Admission for Asthma in Children.* Annals of the American Thoracic Society, 2016. **13**(7): p. 1081-1088.
148. Sharkey, S.W., et al., *Acute and Reversible Cardiomyopathy Provoked by Stress in Women From the United States.* Circulation, 2005. **111**(4): p. 472-479.
149. Hystad, P., et al., *Residential greenness and birth outcomes: evaluating the influence of spatially correlated built-environment factors.* 2014(1552-9924 (Electronic)).

150. Mitchell, R. and F. Popham, *Effect of exposure to natural environment on health inequalities: an observational population study.* The Lancet, 2008. **372**(9650): p. 1655-1660.
151. Brown, S.C., et al., *Neighborhood Greenness and Chronic Health Conditions in Medicare Beneficiaries.* American Journal of Preventive Medicine, 2016. **51**(1): p. 78-89.
152. Wilker, E.H., et al., *Green space and mortality following ischemic stroke.* Environmental Research, 2014. **133**: p. 42-48.
153. Gidlow, C.J., et al., *Where to put your best foot forward: Psycho-physiological responses to walking in natural and urban environments.* Journal of Environmental Psychology, 2016. **45**: p. 22-29.
154. Schnell, I., et al., *Human exposure to environmental health concern by types of urban environment: The case of Tel Aviv.* Environmental Pollution, 2016. **208, Part A**: p. 58-65.
155. Park, B.J., et al., *The physiological effects of Shinrin-yoku (taking in the forest atmosphere or forest bathing): evidence from field experiments in 24 forests across Japan.* Environmental Health and Preventive Medicine, 2009. **15**(1): p. 18-26.
156. Donovan, G.H., et al., *Is tree loss associated with cardiovascular-disease risk in the Women's Health Initiative? A natural experiment.* Health & Place, 2015. **36**: p. 1-7.
157. Augustin, T., et al., *Neighborhood Psychosocial Hazards and Cardiovascular Disease: The Baltimore Memory Study.* American Journal of Public Health, 2008. **98**(9): p. 1664-1670.
158. Barrington, W.E., et al., *Neighborhood socioeconomic deprivation, perceived neighborhood factors, and cortisol responses to induced stress among healthy adults.* Health Place, 2014. **27**: p. 120-6.
159. Foraker, R.E., et al., *Neighborhood socioeconomic disparities and 1-year case fatality after incident myocardial infarction: The Atherosclerosis Risk in Communities (ARIC) Community Surveillance (1992-2002).* American Heart Journal, 2013. **165**(1): p. 102-107.
160. Pickett, K.E. and M. Pearl, *Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review.* Journal of Epidemiology and Community Health, 2001. **55**(2): p. 111-122.
161. Schwartz, B.S., et al., *Body Mass Index and the Built and Social Environments in Children and Adolescents Using Electronic Health Records.* American Journal of Preventive Medicine, 2011. **41**(4): p. e17-e28.
162. Krieger, N., *Theories for social epidemiology in the 21st century: an ecosocial perspective.* International Journal of Epidemiology, 2001. **30**(4): p. 668-677.
163. Steptoe, A. and P.J. Feldman, *Neighborhood problems as sources of chronic stress: Development of a measure of neighborhood problems, and associations with socioeconomic status and health.* Annals of Behavioral Medicine, 2001. **23**(3): p. 177-185.
164. Kwarteng, J.L., et al., *Associations between observed neighborhood characteristics and physical activity: findings from a multiethnic urban community.* Journal of Public Health, 2014. **36**(3): p. 358-367.
165. Lang, I.A., et al., *Neighborhood Deprivation, Individual Socioeconomic Status, and Cognitive Function in Older People: Analyses from the English Longitudinal Study of Ageing.* Journal of the American Geriatrics Society, 2008. **56**(2): p. 191-198.

166. Lindström, M., B.S. Hanson, and P.-O. Östergren, *Socioeconomic differences in leisure-time physical activity: the role of social participation and social capital in shaping health related behaviour.* Social Science & Medicine, 2001. **52**(3): p. 441-451.
167. Matheson, F.I., et al., *Urban neighborhoods, chronic stress, gender and depression.* Social Science & Medicine, 2006. **63**(10): p. 2604-2616.
168. Stockdale, S.E., et al., *The importance of social context: Neighborhood stressors, stress-buffering mechanisms, and alcohol, drug, and mental health disorders.* Social Science & Medicine, 2007. **65**(9): p. 1867-1881.
169. Hawkins, N.M., et al., *Heart failure and socioeconomic status: accumulating evidence of inequality.* Eur J Heart Fail, 2012. **14**(2): p. 138-46.
170. Ramsay, S.E., et al., *Inequalities in heart failure in older men: prospective associations between socioeconomic measures and heart failure incidence in a 10-year follow-up study.* Eur Heart J, 2014. **35**(7): p. 442-7.
171. Kirwa, K., et al., *Residential proximity to major roadways and prevalent hypertension among postmenopausal women: results from the Women's Health Initiative San Diego Cohort.* (2047-9980 (Electronic)).
172. Dadvand, P., et al., *Residential proximity to major roads and term low birth weight: the roles of air pollution, heat, noise, and road-adjacent trees.* (1531-5487 (Electronic)).
173. Zhao, Z., et al., *Residential Proximity to Major Roadways and Risk of Type 2 Diabetes Mellitus: A Meta-Analysis. LID - E3 [pii] LID - 10.3390/ijerph14010003 [doi].* (1660-4601 (Electronic)).
174. DeJarnett, N., et al., *Residential Proximity to Major Roadways Is Associated With Increased Levels of AC133+ Circulating Angiogenic Cells.* (1524-4636 (Electronic)).
175. Allen, R.W., et al., *The spatial relationship between traffic-generated air pollution and noise in 2 US cities.* (1096-0953 (Electronic)).
176. Hart, J.E., et al., *Roadway proximity and risk of sudden cardiac death in women.* (1524-4539 (Electronic)).
177. Kulick, E.R., et al., *Residential Proximity to Major Roadways and Risk of Incident Ischemic Stroke in NOMAS (The Northern Manhattan Study).* (1524-4628 (Electronic)).
178. Alnazzawi, N., et al., *Using text mining techniques to extract phenotypic information from the PhenoCHF corpus.* BMC Medical Informatics and Decision Making, 2015. **15**(2): p. S3.
179. Bielinski, S.J., et al., *A Robust e-Epidemiology Tool in Phenotyping Heart Failure with Differentiation for Preserved and Reduced Ejection Fraction: the Electronic Medical Records and Genomics (eMERGE) Network.* 2015(1937-5395 (Electronic)).
180. Nakamura, A., et al., *Characteristics of heart failure associated with the Great East Japan Earthquake.* (1876-4738 (Electronic)).
181. Salamanca-Bautista, P., et al., *Precipitating factors of heart failure admission: Differences related to age and left ventricular ejection fraction.* International Journal of Cardiology, 2016. **219**: p. 150-155.

Chapter 2: Detailed Methods

2.0 Chapter Overview

The methods underlying this dissertation research center around two main study designs: a nested case-control study of heart failure hospitalization in the cohort of subjects with heart failure in the Geisinger health system (**Chapter 3**) and a cross-sectional study of B-type natriuretic peptide (BNP) levels in blood (**Chapter 4**). Next, I evaluated effect modification by heart failure phenotype (preserved vs. reduced ejection fraction) for both of these analyses (**Chapter 5**). In this chapter, I describe the methods for **Chapters 3** to **5**, including study design, study population, study sample, and analytic variable creation (unconventional natural gas development [UNGD] activity, comorbidities, medication use, environmental variables), with detailed information on preliminary analysis of UNGD activity metrics including evaluation of duration. Only methods that are not explicitly described in **Chapters 3** to **5** are described here.

2.1 Study Population

2.1.1 Geisinger Electronic Health Record (EHR)

Electronic health record (EHR) data were obtained from Geisinger in Pennsylvania. Geisinger is an integrated health system serving over 450,000 primary care patients from over 38 counties in central and Northeastern Pennsylvania who represent the general population in the region [182]. These EHR data are useful for epidemiologic study because they are longitudinal, provide extensive information on a patient's medical history, and are detailed in that they include dates of all encounters for patients within the health care system, including medications, procedures, and laboratory measures [182]. Additionally, the EHR provides accurate and systematic documentation of individual-level factors and comorbidities that are important for

analyses, including age, sex, residential address, race/ethnicity, a measure of poverty (Medical Assistance, a needs-based program, for health insurance), and related risk factors (e.g., BMI, diabetes, chronic obstructive pulmonary disease, cardiovascular conditions such as previous myocardial infarction or coronary artery disease, and other diagnoses) [182].

2.1.2 Selection of Study Sample

The sole criterion for the EHRs we obtained from Geisinger was that the subject had to have at least one *International Classification of Disease* (ICD-9) code for "428.x" for heart failure from any type of encounter. We obtained EHRs for 16,098 subjects who fit these criteria and were seen at a Geisinger facility between 2005-2015. From these 16,098 subjects, we excluded a total of 2915 subjects according to the following: 435 subjects who had either an ICD-9 diagnosis code for congenital heart anomalies (746.x) or endocardial fibroelastosis (425.3); 321 subjects who did not have a residential address in Pennsylvania, or who were unable to be geocoded; two subjects without demographic information; 2071 subjects who were not observed in the EHR between 2008-2015, a time period that coincided with UNGD activity; and 86 subjects who were not at least 18 years old at the start of the study period, January 1, 2008. The remaining 13,183 subjects were eligible for selection into either the hospitalization or BNP analyses.

In **Figure 2.1**, I outline the steps for selecting subjects into the analyses presented in **Chapters 3** to **5**. The rationale and processes through which I selected subjects for the hospitalization analysis and the BNP analysis are described in greater detail in **Chapters 3** and **4**. The study samples for **Chapter 5a** and **Chapter 5b** are the same samples as used in **Chapter 3** and **Chapter 4**, respectively. In **Chapter 5**, I evaluate effect modification of the hospitalization and BNP outcomes by heart failure

phenotypes (comparing subjects with heart failure with preserved [HF*p*EF] vs. reduced ejection fraction [HF*r*EF]) using classifications obtained from the Electronic Medical Records and Genomics Network (eMERGE) phenotype algorithm (explained in next paragraph) [179]. A total of 9639 subjects were selected for either the hospitalization analysis or the BNP analysis, with 9054 of these subjects comprising the study population for the hospitalization analysis (**Chapter 3**) and 3938 subjects included in the BNP analysis (**Chapter 4**). Further detail and rationale for the exclusion criteria and study selection processes can be found in each of the respective chapters. Comparing across these two populations, there were 3353 subjects that were included in both the hospitalization and BNP analyses; 5701 subjects included in the hospitalization analysis that were not included in the BNP analysis; and 585 subjects in the BNP analysis that were not included in the hospitalization analysis. A major reason for the differences in subjects between these two analyses is due to low numbers of BNP orders in later years of the analysis (i.e., 2013-2015), illustrated in **Table 2.1** in section 2.2.2.1 of this chapter. Additional comparisons of the subjects in each of the analyses are described in section 2.7 of this chapter.

After completion of the analyses of the main effects associations of the phase-specific UNGD metrics with heart failure hospitalization status or BNP levels, effect modification by heart failure phenotype on these relations was next evaluated. This required that we use a method to identify heart failure phenotype. Because Geisinger is a member of the eMERGE network, a National Institutes of Health (NIH)-funded consortium of institutions that utilize EHRs and DNA biobanks to develop novel research approaches [183, 184], we decided to use the eMERGE heart failure phenotype method [179]. This method is described in **Chapters 5a** and **5b**, with additional detail in **Appendix A**, but here we describe how application of the phenotype method influenced the samples in the two primary analyses. As **Figure 2.1** outlines, only 5446 of the 13,183

eligible subject records had the eMERGE algorithm successfully applied; this relatively small proportion was due to the abundance of information needed in the EHR to accurately phenotype HF patients with eMERGE, which is detailed in section 2.2.3 of this chapter. Because application of the eMERGE heart failure phenotyping algorithm required additional information, subjects who had the eMERGE algorithm applied had a higher proportion of comorbidities, medications, and had a greater average duration of heart failure (i.e., the time between a subject's first heart failure diagnosis and the date of the case event, control encounter, or laboratory date for BNP) compared to subjects who did not have the algorithm applied (**Tables 2.9-2.13**; further comparisons between study subjects across our analytic samples are detailed in section 2.7 of this chapter). Therefore, findings from **Chapters 5a** and **5b** should be interpreted with caution, since subjects with the eMERGE algorithm applied were likely sicker and certainly had more information in the EHR compared to subjects who did not have the algorithm applied and could therefore bias the findings from **Chapter 5a** and **5b** away from the null.

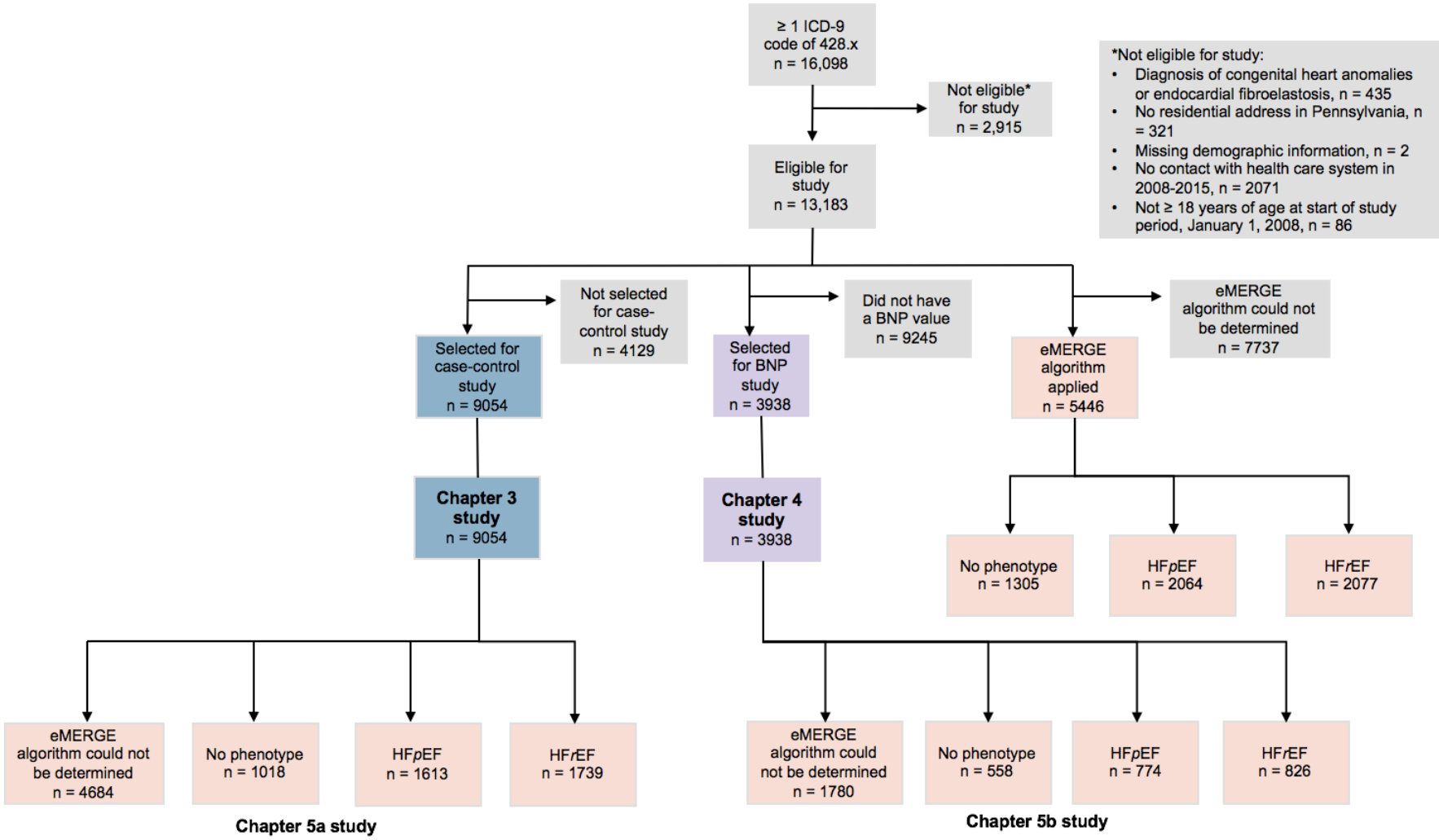


FIGURE 2.2. ALGORITHM FOR STUDY SELECTION

2.2 Study Designs

2.2.1 Hospitalization Analysis

In **Chapter 3** and in **Chapter 5a**, the outcome was hospitalization for heart failure among the previously identified subjects with heart failure.

2.2.1.1 Case Selection

We identified hospitalizations by searching the EHR for a ICD-9 diagnosis code of '428.x' for heart failure associated with two types of encounters, specifically an inpatient encounter and also what the health system refers to as an ED to IPT encounter, in which the subject was first evaluated in the emergency department (ED) and then hospitalized for heart failure. Because some subjects had a number of frequent and recurring hospitalizations, we considered only the subjects' first heart failure hospitalization within the years 2008-2015 as our case events (n = 5839).

2.2.1.2 Control Selection

In **Chapter 3**, I describe in detail the process for control selection. Briefly, we used a 1:1 frequency matching strategy based on the age, sex, and year of the 5839 heart failure subjects' case events. We randomly selected 5839 control encounters, since we needed to assign time-varying covariates and UNGD activity metrics to a date for controls. Control encounters could be any outpatient encounter or medication order, excluding those with a diagnosis code of "428.x" for heart failure. Because of our frequency matching strategy, some subjects were selected for control encounters more than once, leaving us with 3215 control subjects with 5839 control events and a total of 11,678 control encounters and case events for the hospitalization analysis.

2.2.2 BNP Analysis

In **Chapter 4** and in **Chapter 5b**, the outcome evaluated was BNP levels in blood. We identified these laboratory orders and results by searching the EHR laboratory records for procedure codes of '83880.01' and '83880.03' for 'natriuretic peptide' and

'BNP', respectively. We evaluated laboratory BNP values obtained from both inpatient and outpatient settings.

2.2.2.1 NT-PRO-BNP AND BNP: RATIONALE FOR CHOOSING BNP

Increasingly, epidemiologic studies of BNP also evaluate N-terminal pro B-type natriuretic peptide (NT-pro-BNP), which is a non-active prohormone associated with BNP. NT-pro-BNP has a longer half-life than BNP, with NT-pro-BNP having a half-life of approximately 120 minutes compared to a half-life of 20 minutes for BNP [105]. Studies of BNP and NT-pro-BNP show similar prognostic and diagnostic value [100, 105], so we evaluated frequencies of both of these relevant laboratory orders by year and by laboratory setting (**Table 2.1**). Prior to 2012, procedure codes for 83880.01 "Natriuretic peptide" were more prevalent in this population of heart failure patients than procedure codes for 83880.02 "Pro BNP," which became more prevalent in later years in this study. Beginning in 2012, Geisinger began using the procedure code of 83880.03 for "BNP" instead of the previously used procedure code of 83880.01 for "natriuretic peptide." Because of these differences in procedure codes by year, we did not feel it was appropriate to pool the laboratory results from measures of BNP and NT-pro-BNP in a single statistical analysis, we limited our analysis for **Chapter 4** and **Chapter 5b** to laboratory orders for procedure codes of 83880.01 and 83880.03 for measures of BNP. The years of these orders coincided more closely with UNGD activity in the region and were thus more relevant for our evaluation of associations between UNGD activity and BNP values. However, we included measures of BNP after 2012 because clinicians still ordered BNP laboratory tests in later years. To evaluate if the inclusion of BNP beyond 2011 biased our results, we conducted a sensitivity analysis using only 2008-2011 data.

Table 2.1 Summary of all BNP orders, by year, procedure code, and laboratory setting

		Number of laboratory measures (n) Mean concentration (pg/mL) (Standard deviation)		
Year	**Clinical Setting**	**83880.01 "Natriuretic peptide"**	**83880.02 "Pro BNP"**	**83880.03 "BNP"**
2008	All	2,540 546 (759)	262 3233 (6934)	0
	Inpatient	1,410 647 (853)	27 13694 (15,587)	-
	Outpatient	1,130 421 (600)	235 2031 (3544)	-
2009	All	2,718 489 (622)	315 2046 (4252)	0
	Inpatient	1,401 598 (712)	18 5198 (8337)	-
	Outpatient	1,317 374 (483)	297 1855 (3814)	-
2010	All	3,019 514 (654)	381 2256 (3626)	0
	Inpatient	1,560 656 (791)	26 4883 (5797)	-
	Outpatient	1,459 363 (415)	355 2064 (3347)	-
2011	All	3,762 459 (584)	309 2632 (5032)	0
	Inpatient	1,927 578 (704)	35 5868 (7923)	-
	Outpatient	1,835 333 (385)	274 2219 (4386)	-
2012	All	1,199 438 (529)	3,551 4515 (11,635)	175 397 (570)
	Inpatient	602 521 (630)	1740 5913 (8397)	71 492 (679)
	Outpatient	597 354 (383)	1811 3171 (13,932)	104 333 (474)
2013	All	0	5,362 5264 (8519)	482 555 (649)
	Inpatient		2993 6591 (9784)	324 593 (628)
	Outpatient	-	2369 3588 (6190)	158 476 (686)
2014	All	0	5,394 5408 (8880)	284 554 (732)
	Inpatient	-	3167 6723 (10,282)	216 588 (790)
	Outpatient	-	2227 3538 (5894)	68 445 (500)

2015	All	0	2,811 5224 (8341)	4 461 (396)
	Inpatient	-	1688 6467 (9375)	-
	Outpatient	-	1123 3357 (6026)	4 461 (396)

2.2.3 Heart Failure Phenotype Analysis

Analysis for **Chapter 5a** and **Chapter 5b c**lassification of heart failure subjects into the categories of heart failure with preserved vs. reduced ejection fraction (HF*p*EF or HF*r*EF) was determined using the Electronic Medical Records and Genomics (eMERGE) network algorithm for heart failure. This algorithm was developed with funding from the National Human Genome Research Institute, where electronic medical record systems were combined with genetic data from DNA repositories for validation and classification [185]. Briefly, the algorithm for differentiating between HF*p*EF and HF*r*EF phenotypes relies on a collection of diagnoses and procedure codes from the EHR, including physician-interpreted ejection fraction data (both qualitative and quantitative), obtained from echocardiogram records. Necessary information includes: patient demographics, encounter history, ICD-9 diagnosis codes, structured or unstructured problem list, echocardiography measurements, and medications. Because the eMERGE algorithm had these strict case criteria, not all of the subjects in both the case-control study and the BNP study were able to be phenotyped with this algorithm. **Figure 2.2** illustrates directly how the algorithm uses EHR data to differentiate between subjects with HF*p*EF and HF*r*EF, using a cutpoint of an ejection fraction of 50% [185], while also including exclusion criteria, and is also included in **Appendix A. Heart failure algorithm,** which includes detailed information on how the eMERGE algorithm is applied to EHR data for phenotype differentiation. Of note, **Figure 2.2** illustrates that, in order for the phenotyping algorithm to designate a subject as a case, a “HF date assigned per algorithm”; this date is the first date of a heart failure mention in the subject’s problem list

or the first heart failure diagnosis code. One major difference between the criteria for the eMERGE algorithm and the criteria used for selection into the hospitalization or BNP study is that the eMERGE algorithm requires a mention of heart failure in the subject's problem list, whereas we did not consider the problem list for selection of subjects into the hospitalization or BNP study. This is because our sample selection was primarily driven by the presence of a hospitalization for heart failure or a BNP laboratory order, not having heart failure mentioned in the problem list at a previous date. Although we required at least two ICD-9 codes to indicate the presence of comorbidities, we required only a single ICD-9 code for heart failure diagnosis because the positive predictive value and sensitivity of the heart failure ICD-9 diagnosis code 428.x are quite high, with specificity estimated at $\geq$ 95% [186, 187] and estimated positive predictive value ranging from 83.5% to 100% [187, 188]. In the hospitalization analysis, 68.1 % of the 9054 subjects had heart failure mentioned in their problem list; in the BNP analysis, 74.0 % of the 3938 subjects had heart failure mentioned in their problem list. For these reasons and for reasons described in section 2.7 and in **Chapters 5a** and **5b**, we considered both the application of the eMERGE algorithm and the presence of a phenotype for HF*p*EF or HF*r*EF as an indication of more severe disease.

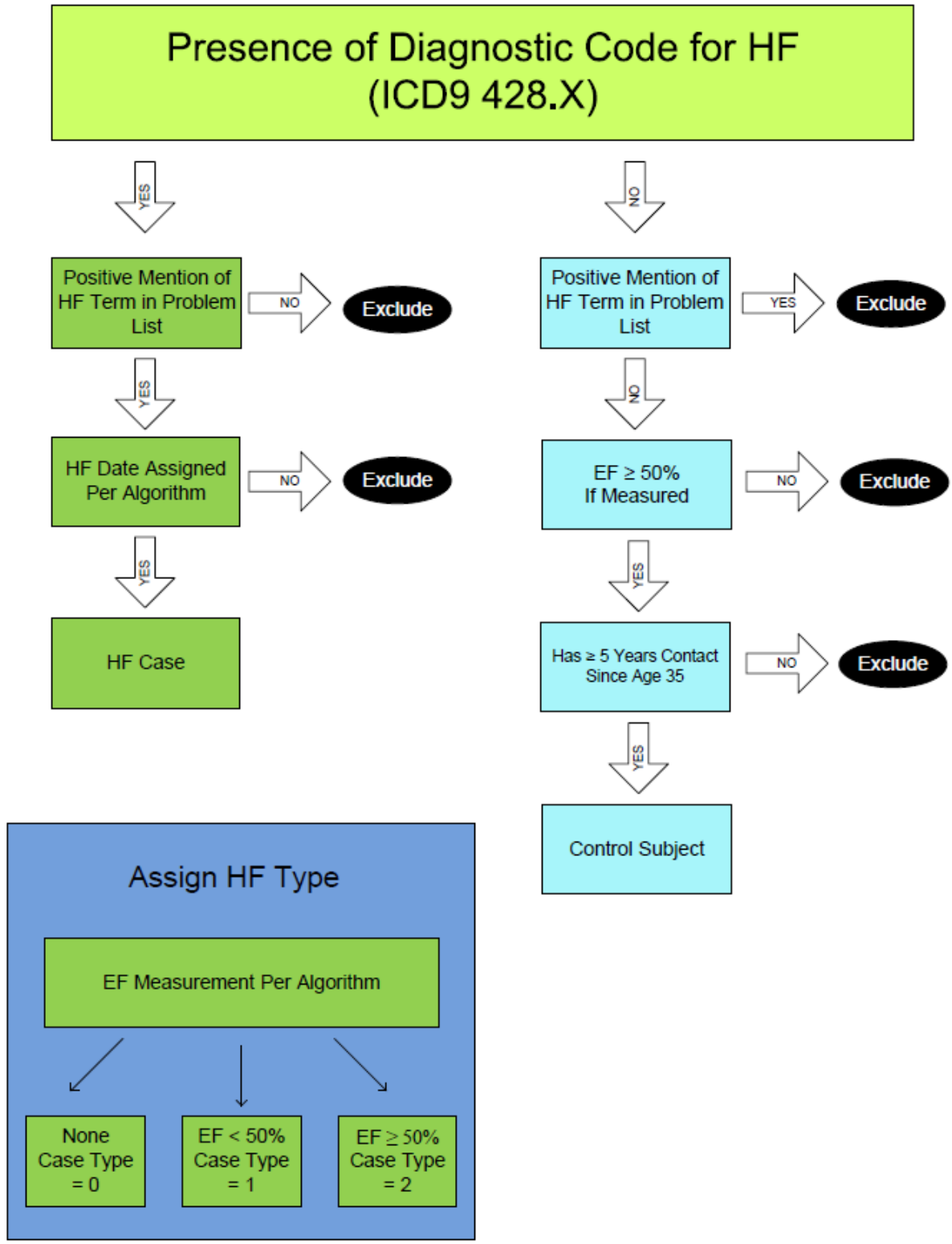


FIGURE 2.3. SCHEMATIC OF THE EMERGE ALGORITHM [185]

2.3 GEOCODING

2.3.1 Patient addresses

Patient addresses were geocoded from address fields located in the EHR. Joseph DeWalle, a geographic information systems (GIS) analyst at Geisinger's Environmental Health Institute, obtained patient address fields from the EHR and processed them for geocoding. Processing of the address fields involved removing extraneous text information and ensuring that address fields were in a correct format (i.e., not a PO box). After processing patient addresses for errors and formatting, addresses were geocoded using ArcGIS v.10. Several basemaps obtained from ESRI (the company that produces ArcGIS software) and the United States Census Bureau's TIGER line shapefiles were used to geocode patient addresses. Latitude and longitude coordinates were obtained using 3-point matching (requiring street address, city, and zip code) and, if city information was unavailable, 2-point matching (requiring street address and zip code), as was previously done in studies in this geography [57, 161, 189]. We ensured that subjects included in this study were able to be geocoded to a street address in Pennsylvania. Of the 16,098 subjects with a diagnosis code of 428.x for heart failure, 16,056 (99.7%) were able to be successfully geocoded. Of the 16,056 geocoded subjects, 15,765 (98.2%) had a residence in Pennsylvania.

2.4.1 EXPOSURE DATA

2.4.1.1 GENERATION OF UNGD ACTIVITY METRICS

Unconventional natural gas development (UNGD) well data were obtained from the Pennsylvania Department of Environmental Protection (PA DEP) for the years 2005-2015. These data included the dates and locations of natural gas wells drilled during this time period. Each natural gas well had its own unique identifier for the well itself and for the well pad on which it was drilled. Data also included the dates when a well was stimulated, and the total volume of natural gas produced over the course of a year. In

2012, the PA DEP began reporting the volume of natural gas produced from a single well in six-month increments instead of in annual increments, as was reported previously. To obtain the total daily volume of natural gas produced by each well, we divided the reported volume by the appropriate number of days for the reporting timeframe. We used these data to identify UNGD activity in four phases: well pad preparation (e.g., clearing of site, delivery of equipment and personnel), beginning of drilling of well (i.e., spud date), stimulation (hydraulic fracturing) of well, and production of natural gas [15].

We generated quantitative estimates of UNGD activity based on the distance of the patient's residential address and the dates of each of the four UNGD phases, as in prior studies of UNGD activity in this region that examined UNGD activity in relation to asthma hospitalizations [15], pregnancy outcomes [13], and nasal and sinus, migraine headache, and fatigue symptoms [17]. We assigned UNGD activity metric values for subjects at a 1-day lag prior to the heart failure hospitalization. For estimating a subject's UNGD activity for the well pad preparation (pad) and spud metrics, we used the following equation developed by Rasmussen et al. [15], where *j* is the subject, *n* is the number of wells and d^2_{ij} is the squared distance (m^2) between *j* and well *i*. subject *j*'s activity metric is estimated by **Equation 1**:

*(**Equation 1**)*

$$\text{Activity metric for patient } j = \sum_{i=1}^{n} \frac{1}{d^2{}_{ij}}$$

Similarly, the stimulation (stim) activity metric for subject *j* was estimated by the

Equation 2:

*(**Equation 2**)*

$$\text{Activity metric for patient } j = \sum_{i=1}^{n} \frac{t_i}{d^2{}_{ij}}$$

where t_i is the total well depth (vertical plus horizontal) and is used as a surrogate for such activities as truck traffic, volume of water brought to site, volume of stimulation chemicals injected into well, volatilization of these chemicals from flowback water into air, drilling engine emissions, and compressor engine emissions (as fluids are injected under high pressure) [15, 190]. The production metric, instead of using total well depth, incorporates the total daily volume of natural gas (m^3) produced for well *i* (v_i), which is used as a surrogate for production activity, compressor engine activity emissions, and fugitive emissions [15, 190]. **Equation 3** was used for the generation of the production (prod) activity metric:

*(**Equation 3**)*

$$\text{Activity metric for patient } j = \sum_{i=1}^{n} \frac{v_i}{d^2_{ij}}$$

In addition to the four phase-specific activity metrics, a composite metric was created as the z-transformed (mean of zero, standard deviation of 1) sum of the four phase-specific metrics for each subject. We evaluated correlations between the composite metric and the phase-specific metrics in section 2.6 of this chapter.

2.5.3 Individual-level covariates

2.5.3.1 Assignment of covariates from the EHR

In **Chapters 3-5**, we assigned a number of time-varying covariates at the individual level to account for comorbid conditions and relevant medications among subjects with heart failure. For comorbidities, we identified these when subjects had least two diagnosis codes in the EHR for each respective comorbid condition prior to the date of the case event or control encounter (**Chapters 3 and 5a**) or prior to the laboratory order date (**Chapters 4 and 5b**). These ICD-9 diagnosis codes could be obtained from any combination of inpatient, outpatient, or emergency department encounters, as well as medication orders, but not from the EHR's problem list. **Table 2.2** lists the ICD-9 codes used to define these key covariates.

Table 2.2. Diagnosis codes used for EHR covariate assignment

Condition	ICD-9 diagnoses codes used
Chronic kidney disease	585.x
Chronic obstructive pulmonary disease	496.x
Coronary artery disease	414.01
Hypertension	401.9
Myocardial infarction	410.x
Type 2 diabetes	250.x
Valve disorders	V43.3, 424.x, 396.9, 391.1, 392.0, 390, 395.9, or 421.9

Relevant medications were assigned in a similar time-varying manner. Medication use was defined as a binary variable (yes vs. no) to indicate whether or not a medication order (which had a defined start and end point) encompassed the date of the case event or control encounter (**Chapters 3 and 5a**) or the laboratory order date (**Chapters 4 and 5b**). Medications were identified using the Medi-Span Generic Product Identified Therapeutic Classification System [191] to sort medications into the following classes: anti-hypertensive, anti-hyperlipidemic, and anticoagulant medications. **Table 2.3** lists the medication subclasses identified by these broader medication class categories.

Table 2.3. Medication classes and subclasses used for covariate assignment

Medication class	Medication subclasses represented
Antihypertensive	Angiotensin converting enzyme (ACE) inhibitors Agents for pheochromocytoma Angiotensin II receptor antagonists Antiadrenergic antihypertensives Antihypertensive combinations Direct renin inhibitors Selective aldosterone receptor antagonists Vasodilators
Antihyperlipidemic	Antihyperlipidemic combinations Bile acid sequestrants Fibric acid derivatives HMG-CoA reductase inhibitors (statins) Intestinal cholesterol absorption inhibitors Miscellaneous antihyperlipidemics Nicotinic acid derivatives
Anticoagulant	Coumarin anticoagulants Direct factor XA inhibitors Heparins and heparinoid-like agents Thrombin inhibitors

We calculated a number of other individual level covariates relevant to analyses in **Chapters 3-5**. These included smoking status, receipt of Medical Assistance, body mass index (BMI), Charlson index, duration of contact with the health care system, and duration of heart failure. Smoking status was calculated from the EHR's social history file and was classified as 'current', 'former', or 'never' at the time of the case or control event or the laboratory order date. Receipt of Medical Assistance was categorized as 'ever' or 'never' having received Medical Assistance based on the presence of insurers listed in **Table 2.4**. Medical Assistance was used as a surrogate for family socioeconomic status as in previous studies [192-194].

Table 2.4. List of insurers considered for receipt of Medical Assistance

Name of insurer
PENNA M A PROGRAM D01
PA MEDICAID PRESUMPTIVE D02
GHP MA G09
CBHNP PA HEALTH CHOICES D04
COVENTRY CARES D06
ACCESS PLUS D15
MEDICAID OF NEW YORK D43
MEDICAID OF OHIO D45
MEDICAID OF MARYLAND D46
MEDICAID OF VIRGINIA D47
MEDICAID SSU D74
MA PENDING D99
MA OUT OF STATE Z43
GATEWAY MDC ASSURD MDC HMO M17
GATEWAY 995
GHP PPO HEALTHCHOICES G22
CHIP UHC COMM PL KIDS H64
GHP CHILD HLTH INS PROG (CHIP)
BLUE CHIP S18
CHIP FIRST PRIORITY HEALTH H31
CHIP CAPITAL BLUE CROSS H20
CHIP HIGHMARK BC AND BS H51
CHIP HIGHMARK BS CENTL PA H53
CHIP KEYSTONE HLTH PL E IBC H55
CHIP FIRST PRIORITY HEALTH H31
CHIP UPMC FOR KIDS I07

BMI was calculated by obtaining the height and weight data for individuals in **Chapters 3** to **5** from the EHR at the closest date prior to, but not on the day of or after, either the control encounter, case event, or laboratory order date. The formula used for calculating this BMI was:

*(**Equation 4**)*

$$\text{BMI (kg / m}^2\text{)} = \text{weight (lbs) X 703} \div \text{height (in}^2\text{)}$$

For individuals without complete height and weight data, we used multiple imputation to estimate BMI based on age at time of case event, control encounter (**Chapters 3 and 5a**), or laboratory date (**Chapters 4 and 5b**), sex, race/ethnicity, smoking status, and receipt of Medical Assistance. To have a composite measure of overall morbidity, we also calculated the Charlson index of comorbidity using 17 items. The 17 items included the following conditions: myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, connective tissue disease, peptic ulcer disease, liver disease (mild and moderate to severe), diabetes (and diabetes with end organ damage), hemiplegia, moderate or severe renal disease, any tumor, leukemia, lymphoma, metastatic tumors, and HIV/AIDS [195, 196]. The Charlson index was calculated at the date of the case event or control encounter (**Chapters 3 and 5b**) or at the laboratory date (**Chapters 4 and 5b**). For individuals without a Charlson index score (some individuals did not have enough contact time with the EHR preceding their event and thus were unable to be assigned a Charlson index score), we imputed Charlson index scores using multiple imputation based on: age, sex, race/ethnicity, smoking status, and receipt of Medical Assistance. Lastly, because we recognized that the duration of contact with the health care system as well as the duration of care for heart failure could affect both outcomes evaluated in **Chapter 3** and **Chapter 4** as either confounding variables or as surrogates for measurement error in variables obtained from the EHR, we created two additional

variables. Duration of contact with the health care system was defined as the elapsed time between an individual's first contact with the health care system (for any condition) and the date of the case event or control encounter (**Chapters 3 and 5b**) or the laboratory date (**Chapters 4 and 5b**). Duration of heart failure was defined as the elapsed time between an individual's first ICD-9 code of '428.x' for heart failure and the date of the case event or control encounter (**Chapters 3 and 5b**) or the laboratory date (**Chapters 4 and 5b**).

2.5.3.2 Individual-level measures of environmental variables

2.5.3.2.1 Distance to major roads, minor roads, and nearest hospital or clinic

We obtained the 2011 and 2013 shapefiles for highways in Pennsylvania and New York from the Federal Highway Administration. These data distinguished roads by the Federal Aid System codes 1-7. These codes were used to designate roads as major roads and minor roads. Major roads included interstates, principal arterial for freeways and expressways, and other principal arterial roads. Minor roads included minor arterials. Other roads that were not used in the major and minor roads designation were major collector roads, urban minor collector roads, and other highways designated as part of the National Highway System. Using ArcGIS and the 'generate near table' function, we calculated the Euclidian distance to major and minor roads from a subject's residential address in meters. Similarly, we calculated the distance to the nearest Geisinger hospital or clinic from a subject's residential address. Geocoded locations of outpatient clinics and hospitals were obtained from the health care system.

2.5.3.2.2 Assignment of the normalized difference vegetation index (NDVI)

We obtained satellite imagery files from the National Aeronautics and Space Administration's (NASA)'s moderate resolution imaging spectroradiometer (MODIS) satellites. NASA operates two MODIS satellites that circulate the globe every 1-2 days: MODIS Terra and MODIS Aqua. These two satellites provide images of the same

surface of the Earth approximately 3 hours apart from each other [197]. NASA derives data in 36 spectral bands (i.e., wavelength groupings) from these two satellites in 16-day increments and in 250m x 250m grids. From these 36 spectral bands, NASA uses the red, near-infrared, and blue wavelengths to derive the normalized difference vegetation index (NDVI), an indicator of green vegetation ranging from 0 to 1, in the 250m x 250m grid [197].

We assigned NDVI to individuals based on the peak greenness (i.e., two-week period with highest NDVI values [maximal greenness]) of each calendar year for the date of the case event or control encounter (**Chapters 3** and **5b**) or the laboratory date (**Chapters 4** and **5b**). Since we used NDVI measures as a proxy for the beneficial effects of greenness exposure, we deemed it appropriate to use the time period where greenness was best captured by satellite imagery, i.e., during the period of peak annual greenness. Subjects were assigned a NDVI value as done previously in this region, by calculating the average NDVI value for an area delineated by a 1250m x 1250m grid that encompasses a patient's residential address [57]. Each subject was assigned a value from 0 to 1, with greater values indicating greater residential greenness.

2.5.4 COMMUNITY-LEVEL COVARIATES

We obtained data from the United States Census Bureau's American Community Survey for the years 2010-2014. Specifically, we used the 2010-2014 ACS 5-year estimates of several community variables at defined geographies (i.e., census tracts). The variables obtained from the ACS were used for the generation of community socioeconomic deprivation (CSD) as previously reported [161, 189, 190, 192, 198]: the percentage of the population with educational attainment less than high school, the percentage of the population not in the labor force, the percentage of the population living below the federal poverty level, the percentage of the population who were

unemployed, the percentage of the population without a car, and the percentage of the population that received public assistance.. Each variable was standardized for direction such that higher values represented more deprivation and was in standard deviation units (after z-transformation) and then summed to create a CSD index value.

Of importance to **Chapters 3-5** is the assignment of community types using a mixed definition of place that is appropriate for the geography of our study area, which encompasses both urban and rural areas. We assigned CSD index values to the community level (i.e., census tracts and minor civil divisions) that encompassed a subject's residential address, with higher values indicating higher levels of deprivation. In cities, we used census tracts; elsewhere within the study area, we used minor civil divisions (MCD, i.e., township, borough, or city) as the spatial unit of analysis for CSD [15, 199]. The mixed definition of place means that we used census tract boundaries for some (i.e., census tracts in cities) and MCD boundaries for others (i.e., townships, boroughs). These methods were consistent for both the case-control and BNP studies (**Chapters 3 and 4**).

2.6 Creation of and rationale for UNGD activity metrics

2.6.1 Evaluation of correlation between metrics

In the case-control study population (**Chapter 3**), we created three different durations of UNGD activity metrics (i.e., 30-day, 60-day, and 90-day durations) with a one-day lag and compared correlations between each of these metrics at the three specified durations. We used a one-day lag for the generation of these UNGD activity metrics because previous studies of the short-term effects of ambient air pollution (one of the suspected mechanisms of UNGD effects on health) and hospitalization for cardiovascular diseases suggest that there is a short latency between exposure and hospitalization [200-203]. In addition to the four phase-specific UNGD activity metrics,

we also generated a 'composite' metric, which was the sum of the z-transformed UNGD activity phases (i.e., pad, spud, stim, and prod) for these three specified durations. We calculated Spearman correlation matrices for each of these metrics for the three specified durations, and we determined that the composite metric did not accurately reflect each of the four phase-specific metrics for any of the durations evaluated. Because the correlations among the four phase-specific metrics using the 30-day duration were in the low to moderate correlation range for some metrics but higher for others, we decided to evaluate each phase-specific metric separately in analyses for **Chapters 3-5**.

2.6.1 Evaluation of different durations for UNGD activity metrics

After deciding that we would use individual, phase-specific UNGD activity metrics, we examined various durations of exposure metrics (i.e., 30-day, 60-day, and 90-day durations) with a one-day lag to determine if differences in duration of the UNGD activity metrics would impact the hospitalization and BNP analyses. We calculated the Spearman correlation coefficients between these different durations for each metric assigned to case events or control encounters in **Chapter 3**. These correlation matrices, displayed in **Tables 2.5-2.8**, showed that the correlations among phase-specific metrics of different durations were highly correlated, with r_s ranging from 0.88 to 1.0. Because of these very high correlations among the various durations, we decided to simplify our approach and use just the 30-day duration phase-specific UNGD activity metrics in the analysis.

Table 2.5. Spearman correlation matrix for pad metrics at 30-, 60-, and 90 - day durations (n = 11,678)

Pad metric	30-day duration, one-day lag	60-day duration, one-day lag	90-day duration, one-day lag
30-day duration, one-day lag	1.00		
60-day duration, one-day lag	0.92	1.00	
90-day duration, one-day lag	0.91	0.99	1.00

Table 2.6. Spearman correlation matrix for spud metrics at 3-0, 60-, and 90- day durations (n = 11,678)

Spud metric	30-day duration, one-day lag	60-day duration, one-day lag	90-day duration, one-day lag
30-day duration, one-day lag	1.00		
60-day duration, one-day lag	0.90	1.00	
90 day duration, one-day lag	0.88	0.98	1.00

Table 2.7. Spearman correlation matrix for stim metrics at 30-, 60-, and 90- day durations (n = 11,678)

Stim metric	30-day duration, one-day lag	60-day duration, one-day lag	90-day duration, one-day lag
30-day duration, one-day lag	1.00		
60-day duration, one-day lag	0.96	1.00	
90-day duration, one-day lag	0.95	0.99	1.00

Table 2.8. Spearman correlation matrix for prod metrics at 30-, 60-, and 90- day durations (n = 11,678)

Prod metric	30-day duration, one-day lag	60-day duration, one-day lag	90-day duration, one-day lag
30-day duration, one-day lag	1.00		
60-day duration, one-day lag	0.9950	1.00	
90-day duration, one-day lag	0.9948	0.9998	1.00

2.7 COMPARISON OF STUDY SUBJECTS BY PHENOTYPE STATUS AND BY STUDY SAMPLE

Because not all subjects in the case control study and in the BNP study were able to have an assigned phenotype from the eMERGE algorithm, I have included several tables comparing subjects who had the eMERGE algorithm applied compared to those who have not. First, **Table 2.9** compares subject characteristics, community type represented, and the mean number of ICD-9 codes for heart failure received for subject across several categories: all of the available 16,098 records, subjects in the hospitalization analysis, subjects in the BNP analysis, subjects in both analyses who had the eMERGE algorithm applied, and, among these, the subjects who had a confirmed HF*p*EF or HF*r*EF phenotype by the eMERGE algorithm. Across all subject categories, the distribution of sex, race/ethnicity, and age was similar, although subjects in the BNP analysis had a mean age of 69.1 years (standard deviation: 12.1 years) and all subjects

in the hospitalization analysis had a mean age of 67.1 years (standard deviation: 12.7 years). The subjects who had the eMERGE algorithm applied and who had a confirmed HF*p*EF or HF*r*EF phenotype had a greater mean number of ICD-9 codes for heart failure from the problem list than did subjects in the hospitalization or BNP analyses. However, subjects in the BNP analysis had a greater mean number of ICD-9 codes for heart failure medications than did subjects in the other groups (mean of 62.8 heart failure ICD-9 codes among the BNP subjects, compared to a mean of 38.6 heart failure medications among subjects in the hospitalization study and a mean of 53.6 heart failure medications among subjects with a HF*p*EF or HF*r*EF phenotype). There was also a greater proportion of subjects who had died in the BNP analysis (48.8% deceased) compared to subjects in the hospitalization analysis (33.7%); the subjects with the eMERGE algorithm applied (35.3%); and the subjects with a HF*p*EF or HF*r*EF phenotype.

Table 2.9. Comparison of subject characteristics across study samples in the hospitalization and BNP analyses and by eMERGE algorithm status

	All subjects with a heart failure diagnosis	Hospitalization analysis subjects	BNP analysis subjects	Subjects with eMERGE algorithm applied	Subjects with HF*p*EF or HF*r*EF phenotype
Total subjects, n = 16098	**n = 16098**	**n = 9054**	**n = 3938**	**n = 5446**	**n = 4141**
Sex, n (%)					
Male	8560 (53.0)	4733 (52.3)	2097 (53.3)	2910 (53.4)	2192 (52.9)
Female	7566 (47.0)	4321 (47.7)	1841 (46.8)	2536 (46.6)	1949 (47.1)
Missing	2 (0)	-	-	-	-
Race/ethnicity, n (%)					
White	15616 (97.0)	8815 (97.4)	3834 (97.4)	5331 (98.0)	4058 (98.0)
Black	221 (1.4)	120 (1.3)	45 (1.1)	52 (1.0)	32 (0.8)
Hispanic	173 (1.07)	88 (1.0)	47 (1.2)	45 (0.8)	35 (0.9)
Other	46 (0.3)	27 (0.3)	11 (0.3)	15 (0.3)	13 (0.3)
Missing	40 (0.3)	4 (0.04)	1 (0.03)	3 (0.06)	3 (0.07)
Age*, years, mean (SD)	67.2 (15.2)	67.1 (12.7)	69.1 (12.1)	68.3 (12.2)	68.0 (12.2)
Missing, n (%)	2 (0.01)	-	-	-	-
Community type, n (%)					
Borough	4885 (30.3)	2825 (31.2)	1287 (32.7)	1716 (31.5)	1311 (31.7)
Township	9123 (56.7)	5174 (57.2)	2196 (55.8)	3208 (58.9)	2435 (58.8)
Census tract (city)	1757 (10.9)	1055 (11.7)	455 (11.6)	522 (9.6)	395 (9.5)
Missing	333 (2.1)	-	-	-	-
Patient status at end of study, n (%)					
Alive	9471 (58.8)	6001 (66.3)	2017 (51.2)	3525 (64.7)	2637 (63.7)
Deceased	6625 (41.2)	3053 (33.7)	1921 (48.8)	1921 (35.3)	1504 (36.3)
Missing	2 (0.01)	-	-	-	-
Number of heart failure ICD-9 codes per subject, mean (SD)					
Problem list	1.3 (1.5)	1.6 (1.7)	1.9 (1.9)	2.1 (1.7)	2.1 (1.7)
Inpatient visits	1.6 (3.2)	2.5 (3.9)	3.4 (5.2)	2.9 (4.5)	3.2 (4.6)
Medications	25.3 (75.6)	38.6 (96.1)	62.8 (131)	51.8 (114.3)	53.6 (117.8)

*Age determined based on subject age at the start of the study at January 1, 2008

Within the context of the hospitalization analysis, **Tables 2.10** and **2.11** illustrate differences in the case control study subjects by whether or not the eMERGE algorithm was able to be applied to these subjects' records. Because the hospitalization analysis utilized time-varying covariates based on the dates of case events and control encounters, these tables are based on a total of 11,678 *events* rather than the 9054 individual subjects, since some subjects served as a control and then later became a case. In **Table 2.10**, we observed that a greater proportion of subjects who had the eMERGE algorithm applied were male; lived in townships and boroughs compared to census tracts; were deceased by the end of the study period (January 31, 2015); and were former smokers. We also observed that the mean age for subjects with the eMERGE algorithm applied was 72.6 years compared to 70.3 years for those who did not have the algorithm applied and that the mean duration, in days, from a case event or control encounter to the subject's first diagnosis of heart failure was slightly greater for those who had the algorithm applied compared to those who did not (1149 days vs. 1056 days). In **Table 2.11**, we display the proportion of subject events in the hospitalization analysis with relevant medications and comorbidities by whether or not the subjects had the eMERGE algorithm applied. Medication use was greater in subjects that had the eMERGE algorithm applied compared to those who did not. Similarly, there was a greater proportion of subjects with the following comorbid diagnoses among subjects with the eMERGE algorithm applied compared to those who did not: chronic obstructive pulmonary disease, coronary artery disease, hypertension, myocardial infarction, valve disorder, type 2 diabetes, chronic kidney disease. Similarly, the mean value for the Charlson index of morbidity was greater (9.23 vs. 8.26) in those with the eMERGE algorithm applied compared to those without.

Table 2.10. Selected subject characteristics, at time of case or control event, among persons in hospitalization analysis, by whether or not the eMERGE algorithm was applied, n = 11,678 events among 9054 persons

Total events n = 11,678	eMERGE not applied n = 5988	eMERGE applied n = 5690
Sex, n (%)		
Male	3076 (51.4)	3072 (54.0)
Female	2912 (48.6)	2618 (46.0)
Race/ethnicity, n (%)		
White	5813 (97.1)	5576 (98.0)
Black	91 (1.5)	56 (1.0)
Hispanic	64 (1.1)	43 (0.8)
Other	17 (0.3)	13 (0.2)
Missing	3 (0.1)	2 (0.04)
Age at hospitalization or at control selection date, years, mean (SD)	70.3 (13.1)	72.6 (12.0)
Age category at first event, n (%) years		
> 18-30	58 (1.0)	26 (0.4)
> 30-40	82 (1.4)	56 (1.0)
> 40-50	295 (4.9)	187 (3.3)
> 50-60	814 (13.6)	624 (11.0)
> 60-70	1435 (24.0)	1179 (20.7)
> 70-80	1660 (27.7)	1748 (30.7)
> 80-90	1569 (26.2)	1783 (31.3)
> 90-100	75 (1.3)	89 (1.6)
Community type, n (%)		
Borough	1835 (30.6)	1805 (31.7)
Township	3373 (56.3)	3346 (58.8)
Census tract (city)	780 (13.0)	539 (9.5)
Community socioeconomic deprivation (CSD),* SD units, quartiles		
1	1091 (18.2)	1092 (19.2)
2	1471 (24.6)	1437 (25.3)
3	1758 (29.9)	1684 (29.6)
4	1668 (27.9)	1477 (26.0)

Total events n = 11,678	**eMERGE not applied n = 5988**	**eMERGE applied n = 5690**
Patient status at end of study, n (%)		
Alive	4295 (71.7)	3625 (63.7)
Deceased	1693 (28.3)	2065 (36.3)
Distance to major road** (meters), mean (SD)	2745 (4184)	2789 (4193)
Distance to minor road** (meters), mean (SD)	1606 (2400)	1561 (2294)
Distance to hospital/clinic (meters), mean (SD)	7005 (8837)	6473 (7584)
Smoking status at event, n (%)		
Current	764 (12.8)	617 (10.8)
Former	2765 (46.2)	2818 (49.5)
Never	2459 (41.1)	2255 (39.6)
Receipt of Medical Assistance† n (%)		
At event	733 (12.2)	605 (10.6)
Body mass index (BMI) at event, kg/m^2, mean (SD)	31.7 (8.3)	31.8 (8.4)
Time since first HF diagnosis, days‡, mean (SD)	1056 (1265)	1149 (1306)

* Community socioeconomic deprivation (CSD) was calculated based on US Census indicators; further information is detailed in the text

**Major & minor roads were identified from the Federal Highway Administration databases; distance from subject's residential address to these roads was calculated in meters

†Medical Assistance, a surrogate for family socioeconomic status, was calculated based on health insurance status at the time of encounters

‡ Days from first HF diagnosis to the date of case or control event

Table 2.11. Selected diagnoses and medication use, at time of case or control event, among persons in hospitalization analysis, by whether or not the eMERGE algorithm was applied, n = 11,678 events among 9054 persons

Total events n = 11678	eMERGE not applied **n = 5988**	eMERGE applied **n = 5690**
Medication use, by class, n (%)*		
Antihypertensive	2509 (41.9)	2941 (54.0)
Antihyperlipidemic	2769 (46.2)	3120 (54.8)
Anticoagulant	1010 (16.9)	1424 (25.0)
Chronic obstructive pulmonary disease (COPD), n (%)	1322 (22.1)	1402 (24.6)
Coronary artery disease (CAD), n (%)	1248 (20.8)	1711 (30.1)
Hypertension, n (%)	4495 (75.1)	4711 (82.8)
Myocardial infarction, n (%)	690 (11.5)	878 (15.4)
Valve disorder, n (%)	1075 (18.0)	1681 (29.5)
Type 2 diabetes, n (%)	2477 (40.9)	2769 (48.7)
Chronic kidney disease, n (%)	1580 (26.4)	2021 (35.5)
Charlson index of morbidity **, mean (SD)	8.26 (3.33)	9.23 (3.31)

* Relevant medication classes were identified based on the dates of physician orders

** A composite measure of overall morbidity; definition described in text

Similar to our evaluation of differences in subject characteristics by whether or not the eMERGE algorithm was applied to subjects in the hospitalization analysis, we also evaluated these differences among subjects in the BNP analysis. To best illustrate these differences by subject, we selected the first BNP measurement per person and displayed characteristics by whether or not the eMERGE algorithm was able to be applied to the subject (**Tables 2.12** and **2.13**). In **Table 2.12**, a slightly higher proportion of subjects with the eMERGE algorithm applied were men, compared to those who did not have the algorithm applied (54.7% vs. 51.5%, $p = 0.04$). Similar to the subjects in the hospitalization analysis, a greater proportion of subjects in the BNP analysis who had the eMERGE algorithm applied lived in Townships and Boroughs compared to census tracts ($p < 0.001$) and had a greater mean duration, in days, from the time of their first diagnosis of heart failure to their laboratory date (921 days vs. 702 days, $p < 0.0001$). However, we did not see as many differences between smoking status, body mass index, receipt of Medical Assistance, or in the proportion of subjects who had died by the end of the study period. **Table 2.13** shows that, a greater proportion of subjects in the BNP analysis who had the eMERGE algorithm applied had used antihypertensive, antihyperlipidemic, and anticoagulant medications ($p < 0.001$, $p < 0.001$, and $p = 0.05$, respectively) compared to those who did not have the algorithm applied. Similar to the subjects in the hospitalization analysis, a greater proportion of subjects in the BNP analysis with the eMERGE algorithm applied had the following comorbidities: chronic obstructive pulmonary disease, coronary artery disease, hypertension, myocardial infarction, valve disorders, type 2 diabetes, and chronic kidney disease ($p < 0.001$ for all of these except for chronic obstructive pulmonary disease, $p = 0.05$). The mean Charlson index of morbidity value was also greater in those with the eMERGE algorithm applied compared to those without ($p = 0.0006$).

Table 2.12. Selected subject characteristics, at the time of each subject's first B-type natriuretic peptide (BNP) laboratory measurement, by whether or not the eMERGE algorithm was applied

Total subjects n = 3938	eMERGE not applied n = 1780	eMERGE applied n = 2158	p-value*
Sex, n (%)			
Male	916 (51.5)	1181 (54.7)	
Female	864 (48.5)	977 (45.3)	0.04
Race/ethnicity, n (%)			
White	1729 (97.1)	2105 (97.5)	
Black	25 (1.4)	20 (0.9)	
Hispanic	21 (1.2)	26 (1.2)	
Other	4 (0.2)	7 (0.3)	
Missing	1 (0.06)	0 (0.0)	0.5
Age at first laboratory date, years, mean (SD)	71.2 (12.5)	71.5 (11.7)	0.3
Community type, n (%)			
Borough	595 (33.4)	692 (32.1)	
Township	935 (52.5)	1261 (58.4)	
Census tract (city)	250 (14.0)	205 (9.5)	< 0.001
Community socioeconomic deprivation (CSD),* SD units, quartiles			
1	440 (24.7)	555 (25.7)	
2	430 (24.2)	559 (25.9)	
3	436 (24.5)	527 (24.4)	
4	474 (26.6)	517 (24.0)	0.2
Patient status at end of study, n (%)			
Alive	915 (51.4)	1102 (51.2)	
Deceased	865 (48.6)	1056 (48.9)	0.8
Distance to major road † (meters), mean (SD)	2481 (3933)	2770 (4244)	0.03
Distance to minor road † (meters), mean (SD)	1475 (2309)	1572 (2330)	0.2
Distance to hospital/clinic (meters), mean (SD)	6607 (8897)	6423 (7674)	0.5

Total subjects n = 3938	eMERGE not applied n = 1780	eMERGE applied n = 2158	p-value*
Smoking status at event, n (%)			
Current	199 (11.2)	241 (11.2)	
Former	856 (48.1)	1055 (48.9)	
Never	725 (40.7)	862 (39.9)	0.9
Receipt of Medical Assistance‡ n (%)	170 (9.6)	238 (11.0)	0.1
Body mass index (BMI) at event, kg/m^2, mean (SD)	31.9 (7.5)	32.3 (7.3)	0.1
Time since first HF diagnosis, days§, mean (SD)	702 (992)	921 (1104)	< 0.0001

*p-value calculated by chi2 for categorical variables and analysis of variance (ANOVA) F-test for continuous variables

* Community socioeconomic deprivation (CSD) was calculated based on US Census indicators; further information is detailed in the text

†Major & minor roads were identified from the Federal Highway Administration databases; distance from subject's residential address to these roads was calculated in meters

‡ Medical Assistance, a surrogate for family socioeconomic status, was calculated based on health insurance status at the time of encounters

§ Days from first heart failure diagnosis to the date of case or control event

Table 2.13. Selected diagnoses and medication use at time of first B-type natriuretic peptide (BNP) laboratory order, by heart failure phenotype status

Total subjects n = 3938	eMERGE not applied **n = 1780**	eMERGE applied **n = 2158**	p-value*
Medication use, by class, n (%)			
Antihypertensive	699 (39.3)	986 (45.7)	< 0.001
Antihyperlipidemic	778 (43.7)	1068 (49.5)	< 0.001
Anticoagulant	360 (20.2)	492 (22.8)	0.05
Chronic obstructive pulmonary disease (COPD), n (%)	375 (21.1)	474 (22.0)	0.5
Coronary artery disease (CAD), n (%)	272 (15.3)	532 (24.7)	< 0.001
Hypertension, n (%)	1090 (61.2)	1481 (68.6)	< 0.001
Myocardial infarction, n (%)	126 (7.1)	234 (10.8)	< 0.001
Valve disorder, n (%)	254 (14.3)	497 (23.0)	< 0.001
Type 2 diabetes, n (%)	655 (36.8)	922 (42.7)	< 0.001
Chronic kidney disease, n (%)	706 (39.7)	975 (45.2)	< 0.001
Charlson index of morbidity †, mean (SD)	8.6 (3.1)	8.9 (2.8)	0.0006

* p-value obtained from either chi^2 tests (for categorical variables) or analysis of variance (ANOVA) F-test (for continuous variables), comparing subjects who had the phenotyping algorithm (eMERGE) applied vs. those who did not have the necessary information to have this algorithm applied

In both the hospitalization and BNP analyses, it is clear that subjects who had the eMERGE algorithm applied tended to have more comorbidities, medications, and a longer duration of heart failure (as calculated from the EHR) (**Tables 2.10-2.13**). In both of these analyses, a smaller proportion of subjects who had the eMERGE algorithm applied were female and lived in census tracts (cities). It is possible that analyses reliant on the eMERGE algorithm (i.e., **Chapter 5a** and **Chapter 5b** analyses) could be biased away from the null because of these differences, which suggested that those who had the eMERGE algorithm applied had more frequent utilization of health care services. However, in **Chapter 5**, I explore these differences in greater detail and by subjects who had a confirmed HF*p*EF or HF*r*EF phenotype, within the context of the hospitalization and the BNP analyses. Additionally, I evaluated all of the variables (e.g., comorbidities, sex, duration of heart failure) that differed between subjects who had the eMERGE algorithm applied and subjects who did not have the eMERGE algorithm applied in **Chapters 5a** and **5b**.

2.8 Data analysis

The primary descriptions of the statistical analyses are in the relevant specific chapters. Here I describe additional details not included in those chapters.

2.8.1 Univariate analysis and bivariate associations

For all variables considered in **Chapters 3-5**, we examined distributions and measures of central tendency, range, and dispersion across all available data; across study subjects (often using one randomly selected observation per person to evaluate the distributions of time-varying variables across study subjects); and within individual subjects over time, where longitudinal data were available. We used scatterplots and Spearman's correlation coefficients to examine bivariate associations for continuous variables; stratified histogram plots for examining bivariate associations between

categorical and continuous variables; and frequency tables with chi^2 tests for evaluating bivariate associations between categorical variables.

2.8.2 Inverse probability weighting to account for selection bias in hospitalization and BNP study

In **Figure 2.1**, I note that 13,183 individuals were "eligible for selection" into the analyses presented in **Chapters 3-5**. To account for any potential selection bias in analyses presented in **Chapter 4** compared to **Chapter 3**, I generated inverse probability weights from logistic regression models that predicted the probability of being selected in the BNP analysis (**Chapter 4**). These logistic regression models adjusted for smoking status (ever vs. never), total contact time with the EHR (centered and centered-squared), Charlson index (centered and centered-squared), age at end of study (centered), diagnoses of: myocardial infarction, coronary artery disease, valve disorder, chronic kidney disease, and ever receiving anticoagulant medication. The Pearson's goodness of fit test for this model yielded a p-value of 0.33 and the Hosmer-Lemeshow goodneww of fit test, using 10 groups, yielded a p-value of 0.75, indicating that the model adequately fit these data. We calculated the predicted probabilities of inclusion in the BNP analysis from this model (described in greater detail in **Chapter 4**), and we used the inverse of these probabilities in a weighted analysis accounting for the probability of inclusion. The results of sensitivity analyses that account for these inverse-probability weights in the hospitalization and BNP analyses are included in **Chapter 4**.

2.9 REFERENCES

1. Casey, J.A., et al., *Using Electronic Health Records for Population Health Research: A Review of Methods and Applications.* Annual Review of Public Health, 2016. **37**(1): p. 61-81.
2. Bielinski, S.J., et al., *A Robust e-Epidemiology Tool in Phenotyping Heart Failure with Differentiation for Preserved and Reduced Ejection Fraction: the Electronic Medical Records and Genomics (eMERGE) Network.* 2015(1937-5395 (Electronic)).
3. Institute, G.M. *eMERGE Electronic Medical Records and Genomics Network*. 2019 [cited 2019; Available from: https://www.geisinger.edu/research/departments-and-centers/gmi/emerge-electronic-medical-records-and-genomics-network.
4. Health;, N.I.o. *Electronic Medical Records and Genomics Network (eMERGE)*. 2019; Available from: https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE.
5. Weber, M., C. Mitrovic V Fau - Hamm, and C. Hamm, *B-type natriuretic peptide and N-terminal pro-B-type natriuretic peptide - Diagnostic role in stable coronary artery disease.* (1205-6626 (Print)).
6. Maries, L. and I. Manitiu, *Diagnostic and prognostic values of B-type natriuretic peptides (BNP) and N-terminal fragment brain natriuretic peptides (NT-pro-BNP).* (1680-0745 (Electronic)).
7. Bielinski;, S.J., *Heart Failure (HF) with Differentiation between Preserved and Reduced Ejection Fraction*, in *PheKB*, M. Clinic., Editor. 2013.
8. Rosamond Wayne, D., et al., *Classification of Heart Failure in the Atherosclerosis Risk in Communities (ARIC) Study.* Circulation: Heart Failure, 2012. **5**(2): p. 152-159.
9. Goff, D.C., Jr., et al., *Congestive Heart Failure in the United States: Is There More Than Meets the I(CD Code)? The Corpus Christi Heart Project.* JAMA Internal Medicine, 2000. **160**(2): p. 197-202.
10. Grijalva, C.G., et al., *Computerized definitions showed high positive predictive values for identifying hospitalizations for congestive heart failure and selected infections in Medicaid enrollees with rheumatoid arthritis.* Pharmacoepidemiology and drug safety, 2008. **17**(9): p. 890-895.
11. Casey, J.A., et al., *Greenness and Birth Outcomes in a Range of Pennsylvania Communities.* International Journal of Environmental Research and Public Health, 2016. **13**(3).
12. Schwartz, B.S., et al., *Body Mass Index and the Built and Social Environments in Children and Adolescents Using Electronic Health Records.* American Journal of Preventive Medicine, 2011. **41**(4): p. e17-e28.
13. Liu, A.Y., et al., *The contextual influence of coal abandoned mine lands in communities and type 2 diabetes in Pennsylvania.* Health & Place, 2013. **22**: p. 115-122.
14. Rasmussen, S.G., et al., *Association Between Unconventional Natural Gas Development in the Marcellus Shale and Asthma Exacerbations. .* JAMA Internal Medicine, 2016(2168-6114).
15. Casey, J.A., et al., *Unconventional Natural Gas Development and Birth Outcomes in Pennsylvania, USA.* Epidemiology, 2015.
16. Tustin, A.W., et al., *Associations between Unconventional Natural Gas Development and Nasal and Sinus, Migraine Headache, and Fatigue Symptoms in Pennsylvania.* Environ Health Perspect, 2016.

17. Koehler, K., et al., *Exposure Assessment Using Secondary Data Sources in Unconventional Natural Gas Development and Health Studies.* Environmental Science & Technology, 2018. **52**(10): p. 6061-6069.
18. Medi-Span, *Master Drug Data Base Documentation Manual*. 2007, Indianapolis, Indiana.
19. Casey, J.A., et al., *Measures of SES for Electronic Health Record-based Research.* American Journal of Preventive Medicine, 2018. **54**(3): p. 430-439.
20. Hirsch, A.G., et al., *Early-life antibiotic use and subsequent diagnosis of food allergy and allergic diseases.* Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology, 2017. **47**(2): p. 236-244.
21. Schwartz, B.S., et al., *Attention deficit disorder, stimulant use, and childhood body mass index trajectory.* Pediatrics, 2014. **133**(4): p. 668-676.
22. Charlson, M., et al., *A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.* (0021-9681 (Print)).
23. Chuang, J.H., C. Friedman, and G. Hripcsak, *A comparison of the Charlson comorbidities derived from medical language processing and administrative data*, in *Proc AMIA Symp*. 2002.
24. NASA;. *NASA Distributed Active Archive Center (DAAC); MODIS Data: Terra vs. Aqua*. 2019; Available from: https://nsidc.org/data/modis/terra_aqua_differences.
25. *American Community Survey*. 2010, United States Census Bureau.
26. Schwartz, B.S., et al., *Body mass index and the built and social environments in children and adolescents using electronic health records.* (1873-2607 (Electronic)).
27. Martins, L.C., et al., *The effects of air pollution on cardiovascular diseases: lag structures.* (0034-8910 (Print)).
28. Braga, A.L., J. Zanobetti A Fau - Schwartz, and J. Schwartz, *The lag structure between particulate air pollution and respiratory and cardiovascular deaths in 10 US cities.* (1076-2752 (Print)).
29. Guo, B., et al., *Using spatio-temporal lagged association pattern to unravel the acute effect of air pollution on mortality.* (1879-1026 (Electronic)).
30. de Paula Santos, U., et al., *Effects of air pollution on blood pressure and heart rate variability: a panel study of vehicular traffic controllers in the city of Sao Paulo, Brazil.* (0195-668X (Print)).

Chapter 3: Association between unconventional natural gas development activity and hospitalization among patients with heart failure in Pennsylvania, 2008-2015.

Authors: Tara P. McAlexander, MPH[1], Karen Bandeen-Roche, PhD, MS[2,3], Jessie P. Buckley, PhD, MPH[1,4], Jonathan Pollak, MS[1], Erin D. Michos, MD, MHS[3], John William McEvoy, MBBCh, MHS[3,5], Brian S. Schwartz, MD, MS[1,3,4]

Affiliations
[1] Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States
[2] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States
[3] Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States
[4] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States
[5] National Institute for Preventive Cardiology, National University of Ireland, Galway, Ireland

Corresponding Author: Brian S. Schwartz, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Room W7041, Baltimore, Maryland 21205, email bschwar1@jhu.edu
Running title: UNGD and heart failure hospitalization

Acknowledgements
This research was supported by funding from the National Institute of Environmental Health Sciences (NIEHS) training grant # 5T32ES007141-32 (T McAlexander) and NIEHS grant ES023675-01 (PI: B Schwartz). Additional support was provided by the Degenstein foundation for the compiling of well data. The authors would like to thank Joseph DeWalle, Dione Mercer, and staff of the Geisinger Environmental Health Institute for their assistance in obtaining data essential for this study.

Competing financial interest disclosure: The authors declare that they have no competing financial interests.

3.1 ABSTRACT

Background: A growing number of epidemiologic studies report associations between unconventional natural gas development (UNGD) activity and several health outcomes; hypothesized underlying pathways include air pollution, noise and stress. Heart failure subjects are susceptible to environmental stressors and hospitalization for exacerbations, making it a logical outcome for investigation.

Objectives: To evaluate the association between measures of UNGD activity and the odds of hospitalization among heart failure subjects living in Pennsylvania, a state with active UNGD in the Marcellus shale.

Methods: We conducted a nested case-control study of hospitalization among persons with heart failure seen at Geisinger from 2008 to 2015 using electronic health record (EHR) data. We frequency-matched persons with heart failure and hospitalizations to randomly selected control encounters among heart failure subjects without hospitalizations, by age, sex, and year. We assigned metrics of UNGD activity by phase in the 30 days prior to hospitalization or control encounter and compared the odds of hospitalization by quartile, adjusting for confounding variables with particular attention to potential spatial and temporal confounding.

Results: We identified 9,054 heart failure subjects, 47.7% (n = 4,321) of whom were female, with a mean (SD) age of 71.1 (SD) years and 5,839 hospitalizations for heart failure. Comparing the 4^{th} to 1^{st} quartiles of UNGD activity, we found associations with hospitalization (OR [95% CI]) of 1.70 (1.35 - 2.13), 0.97 (0.75-1.27), 1.80 (1.35-2.40), and 1.62 (1.07-2.45) for the phases of pad preparation, well drilling, well stimulation, and natural gas production, respectively. Several metrics evidenced exposure-effect relations across UNGD quartiles.

Conclusions: This is the first study to directly assess hospitalization among heart failure subjects in relation to UNGD in a large patient population. Most prior studies of UNGD

and health focused on health conditions that mainly affected young and middle-aged persons; heart failure is a condition that mainly affects older persons, who should be more susceptible to the exposures that arise from UNGD, Three of four phases of UNGD activity were associated with increased odds of hospitalization, findings which are both biologically plausible and consistent with a growing body of epidemiological evidence suggesting negative health effects are associated with UNGD activity.

3.2 INTRODUCTION

Heart failure is a common chronic condition which effects over 5.7 million Americans and 25 million persons globally [73-75, 204]. Known risk factors for heart failure include coronary artery disease, hypertension, ischemic heart disease, atrial fibrillation, diabetes [86, 89], chronic obstructive pulmonary disease [205], depression and anxiety [143, 206, 207], and lifestyle factors that both directly and indirectly worsen heart failure including tobacco smoking, lack of physical activity, and poor diet [73, 208]. Collectively, heart failure costs the US health care system over $30 billion in direct (e.g., pharmaceuticals, health care services) and indirect (e.g., missed work days) costs annually [20, 21, 73].

Clinically, there are two main phenotypes of heart failure: heart failure with reduced ejection fraction (HF*r*EF) and heart failure with preserved ejection fraction (HF*p*EF) [209, 210]. In both phenotypes, heart failure subjects have impairment of the left ventricle's pump capacity, resulting in reduced blood flow to critical organs (e.g., kidneys, lungs, brain) [211]. Symptoms of acute heart failure include but are not limited to: dyspnea, fatigue, irregular heartbeat, persistent cough, and fluid retention, swelling of the legs, and rapid weight gain due to reduced blood flow to these vital organs [212]. Given these symptoms, subjects with heart failure are at risk for frequent hospitalizations and mortality due to worsening of left ventricular pump function and the co-occurrence of

related conditions (e.g., hypertension, diabetes, coronary artery disease) [89]. Understanding the full spectrum of risk factors associated with these severe and acute symptoms is thus vitally important for improved patient outcomes and quality of life.

Persons living with heart failure are also susceptible to environmental exposures [213]. A growing number of epidemiologic studies have found associations between environmental factors, particularly air pollution (i.e., $PM_{2.5}$), and hospital admissions for heart failure [26, 130, 214, 215]. Suspected biologic mechanisms underlying these associations include systemic inflammation, direct tissue injury, ischemia, arrhythmias, and thrombosis [213, 216]. There is also a growing body of literature linking environmental noise exposure to the worsening of hypertension, coronary artery disease, and heart rate variability [217, 218]. Psychosocial stress has also been linked to coronary artery disease and hypertension [143, 219], two important risk factors for heart failure.

Unconventional natural gas development (UNGD) is a growing industry worldwide, and Pennsylvania, US has seen substantial growth, where over 9000 wells have been drilled in the Marcellus shale since 2004 [61]. UNGD has a number of environmental impacts such as increased noise and air pollution levels (e.g., $PM_{2.5}$, oxides of nitrogen [NO_x], oxides of sulfur [SO_x], volatile organic compounds [VOCs], polycyclic aromatic hydrocarbons [PAHs]) associated with its several stages [22-25]. A growing literature has identified associations between metrics of UNGD activity and health outcomes including low birth weight, small for gestational age [9, 10]; preterm birth [11-13]; congenital defects [14]; three types of asthma exacerbations [15, 16]; and migraine, fatigue, nasal and sinus symptoms (Tustin, Hirsch et al. 2016), and depression symptoms [18]. None of these studies have concurrently measured exposure elements to identify the likely impact pathways, but the existing health literature clearly documents

the biological plausibility of these health impacts through several candidate mechanisms [18, 27, 66].

The purpose of this study was to evaluate associations between UNGD activity metrics, by phase of development, and hospitalization among subjects with heart failure. To the best of our knowledge, no prior epidemiologic studies have examined associations between measures of UNGD activity and heart failure outcomes in a large subject population.

3.3 Methods and Measures

3.3.1 Study population and study design

We conducted a case-control study, comparing persons with heart failure with and without hospitalizations, using electronic health record (EHR) data from Geisinger, an integrated health system with multiple inpatient and outpatient centers in Pennsylvania, for January 1, 2005 to July 31, 2015. The study was nested within the general-population-representative, open, dynamic cohort that persons with a Geisinger primary care provider represent [182]. We searched patient records for at least one *International Classification of Disease* (ICD-9) diagnosis code of 428.x for heart failure from inpatient, outpatient, or emergency department (ED) encounters and medication records, excluding those from laboratory orders and the EHR problem list. We identified 16,098 subjects with a heart failure ICD-9 code. From these 16,098 subjects, we excluded 435 individuals with ICD-9 codes for congenital heart anomalies (n = 424) and endocardial fibroelastosis (n = 14) (746.x or 425.3, respectively). We also excluded 321 individuals who did not reside in Pennsylvania or did not have geocoded residential information, and 2 individuals with missing demographic information. The analysis was limited to 2008 to 2015, resulting in exclusion of 2071 more persons with heart failure, to coincide with the onset of UNGD activity in Pennsylvania, resulting in 13,183 subjects.

We identified all events for these subjects and excluded all subjects who were not at least 18 years of age at the date of both the event and at time of heart failure diagnosis (i.e., their first ICD-9 code of “428.x”), leaving us with 13,183 subjects eligible. Lastly, we excluded 853 subjects who did not have an ICD-9 code for 428.x in 2008-2015. After all exclusions, there were 12,330 individuals who were eligible for selection in the case-control study (**Figure 3.1**).

3.3.2 Case identification

From the pool of 12,330 eligible subjects with heart failure, we identified 5,839 subjects who were hospitalized for heart failure during the study period (i.e., these subjects had the 428.x ICD code associated with an inpatient encounter or an encounter designated as emergency to inpatient). We included only incident heart failure hospitalizations, i.e., only the first hospitalization recorded within the study period, in this analysis.

3.3.3 Control selection

Subjects with heart failure were eligible for control selection if they had not been hospitalized for heart failure up to 30 days before their randomly selected encounter date in the year of the case’s heart failure hospitalization. To limit the potential for confounding by age, sex, and calendar year, we used incidence density sampling with replacement of selected controls and frequency-matched cases to control encounters by these variables [220]. Age was categorized for frequency-matching as: > 18–30 years, > 30–40 years, > 40–50 years, > 50–60 years, > 60–70 years, > 70–80 years, > 80–90 years, and > 90–100 years. Subjects who did not yet or did not ever have a hospitalization for heart failure were eligible for control selection encounters, which included any outpatient visit or medication order that did not include a heart failure

diagnosis; a subject was able to be selected only once per calendar year for a maximum of five control encounters. Other aspects of timing were not used in the control selection process, however we evaluated EHR contact time as well as duration of heart failure diagnosis in our model building process. Using this frequency-matching strategy, we randomly selected 5839 control encounters in the year of hospitalization for cases because time-varying covariates required a date for calculation. In sensitivity analyses, analysis was repeated using a 4:1 control to case matching ratio to optimize power; this was not the primary analysis because, in order to match at a 4:1 ratio according to age category, sex, and year, control subjects in the tails of the age distribution were used multiple times, and we wanted to limit any undue influence from utilizing some controls multiple times in our primary analysis.

3.3.4 COVARIATE ASSIGNMENT

Information for time-invariant subject characteristics, such as sex and race/ethnicity, were available from the EHR demographics file. We calculated a number of time-varying covariates for all subjects in the analysis, including age at hospitalization or control encounter date, smoking status (never, previous, current), Charlson index of morbidity [221], and receipt of Medical Assistance (a surrogate for family socioeconomic status) as previously reported [192, 222]. We also identified the presence of co-morbid conditions based on at least two encounter diagnosis codes (ICD-9) on any date between January 1, 2008 and the day of the hospitalization or control encounter date, including type 2 diabetes (250.x), hypertension (401.9), chronic obstructive pulmonary disease (496.x), coronary artery disease (414.01), valve disorders (V43.3, 424.x, 396.9, 391.1, 392.0, 390, 395.9, or 421.9), previous myocardial infarction (410.x), or diagnosis of chronic kidney disease (585.x).

We calculated the duration (days) from subjects' first contact with Geisinger to the date of the hospitalization or encounter date. Season of hospitalization or control encounter was defined as: winter (December 22 – March 20); spring (March 21 – June 20); summer (June 21 – September 21); and fall (September 22 – December 21). We used medication records to identify current medication use by verifying that the date of hospitalization or control encounter date was between the start and end dates of the medication order. Current medication use was identified for the following classes of medications: antihypertensive, antihyperlidemic, and anticoagulant, all of which could impact a subject's experience of symptoms, exacerbation status, and likelihood of hospitalization [223-225].

We used available height and weight measurements to calculate body mass index (BMI, kg/m^2) at the date closest to, but no more than 365 days prior to, either the heart failure hospitalization date or control encounter date. Only biologically plausible values were used (height from 36 to 90 inches, weight from 50 to 600 pounds). For individuals without sufficient height and weight data to calculate BMI within one year of the case event or control encounter date (n = 339, or approximately 3.7% of the 9054 study subjects), we imputed BMI using multiple imputation based on age, sex, and receipt of Medical Assistance [226].

3.3.5 Assignment of community metrics

We used subjects' residential addresses from the EHR to obtain latitude and longitude coordinates as previously reported [13, 190, 192]. These coordinates were used to identify the community type of the residential location as township, borough, or city, first using minor civil division shape files, and then further subdivided residential locations in cities to the census tract using census tract shape files, as previously reported [161, 189, 198]. Residential locations were also grouped into five sub-regions of

our 38-county study area consisting of 5 to 15 contiguous counties each, identified as northeast, southeast, central, southwest, and northwest, to account for potential spatial confounding at a scale larger than the level of community type. This variable was created after visualizing assigned UNGD activity metric values across the study area and observing different spatial distributions comparing the spud metric (i.e., drilling activity) to the production metric (i.e., cumulative natural gas production). For townships, boroughs, and census tracts in cities we calculated community socioeconomic deprivation using 2010-2014 data from the US Census American Community Survey as previously reported [161, 189, 190, 198]. Community socioeconomic deprivation was standardized for direction such that higher values represented more deprivation and was in standard deviation units (after z-transformation of its components before summing). CSD was categorized into quartiles for analysis based on the entire study population.

We also obtained the locations of major (i.e., highways) and minor roads (i.e., arterial and local roads) from the Federal Highway Administration. Using ArcGIS (ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute), we calculated the Euclidian distance from each subject's residential address to both major and minor roads separately, in meters. In the same manner, we also obtained the geographic locations of all Geisinger facilities, and we calculated, in meters, the distance from each subject's residential address to the location of the nearest hospital or clinic.

3.3.6 UNGD ACTIVITY ASSIGNMENT

UNGD activity metric data and calculation have been previously reported [190]. In brief, we obtained data from the Pennsylvania Department of Environmental Protection on UNGD wells for the years 2005 to 2015, documenting the dates and locations of four phases of UNGD activity: well pad preparation (e.g., clearing of site,

delivery of equipment and personnel), drilling of well (i.e., starting at the spud date), stimulation (hydraulic fracturing) of well, and production of natural gas. We assigned UNGD activity metrics using R (R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria) that incorporated number, phase, size, and location of wells, and divided by the squared distance from residential locations to all wells in the state, as previously reported [13, 15, 17, 62, 190]. UNGD activity assignments used **Equation 1** [15] where *j* identified subject, *n* was the number of wells, and d^2_{ij} was the squared distance (m^2) between subject *j* 's residential address and well *i*.

(Equation 1) Activity metric for subject *j* $= \sum_{i=1}^{n} \frac{m_i}{d^2{}_{ij}}$

For the pad preparation and spud activity metric, m_i = 1. For the stimulation and production metrics, m_i was total well depth (m) or total daily volume of natural gas (m^3) produced for well *i,* respectively [15]. We calculated each of these activity metrics for a duration of 30 days before the hospitalization or control encounter date with a one day lag (i.e., activity was not counted in the day before, **Figure 3.2**). A one-day lag was chosen because we hypothesized that there would be a lag between UNGD activity and the worsening of heart failure symptoms that would lead a subject to seek care and become hospitalized, which is consistent with many studies of air pollution and respiratory and cardiovascular outcomes that also use a relatively short exposure period with a 1-day lag prior to hospitalization [26, 142, 227]. We also examined correlations between various durations to understand the extent to which these UNGD activity metrics were correlated (see **Chapter 2, Section 2.5**).

3.3.7 STATISTICAL METHODS

We compared the distribution of individual-level covariates, community metrics (i.e., community socioeconomic deprivation, distance to roads and hospitals), and UNGD

activity metrics by case and control subjects. Because our study population resided throughout Pennsylvania, and because our study spanned 7.5 years, we examined the distributions of UNGD activity metrics by phase, by quartile of each phase, and by year to understand any temporal patterns in the distribution of quartiles of UNGD activity metrics. Similarly, we examined the frequencies of subjects in each quartile of the four phases of UNGD activity by heart failure hospitalization status (i.e., comparing subjects who were hospitalized for heart failure at any point to those who were never hospitalized for heart failure) to understand the crude associations before developing a model for heart failure hospitalization.

We used the *melogit* function in Stata v13.1 (StataCorp LP 2016. Stata/MP 13.1. College Station, TX.) to develop multi-level logistic regression models that estimated the odds of hospitalization, comparing cases to controls, by quartile of UNGD activity. We included random intercepts for subject (to account for correlation within individuals over time who were included in analysis more than once as control then as case) and community type (to account for the correlation of measures for persons clustered in communities). We evaluated non-linearity for continuous variables (e.g., BMI, duration of heart failure, and distance to road measures) by evaluating linear, quadratic, and cubic terms after centering of the variable; higher order terms were only included if the association crossed an inferential threshold ($p < 0.05$). Age was included in models as the categorized frequency-matching variable, although we did evaluate age as a linear, quadratic and cubic term after centering, in the same manner as we evaluated other linear variables.

Our initial model (**Table 3.6, Model 1**) included the following variables: sex (female vs. male), age category at hospitalization or control encounter date (as previously described), as well as race/ethnicity (nonwhite vs. white), BMI, and Medical Assistance based on prior evidence that these are strongly associated with

hospitalization for heart failure [228]. Additional models evaluated, in a stepwise fashion, the inclusion of year of hospitalization or control encounter (**Table 3.6, Model 2**); geographic region (**Table 3.6, Model 3**); both year and region (**Table 3.6, Model 4**); and lastly, the additional inclusion of season, EHR contact time (the date of a subject's hospitalization or control encounter minus the date of the subject's first observation in the EHR), and distance to nearest hospital or clinic (**Table 3.6, Model 5**). We also evaluated models that additionally adjusted for distance to both major and minor roads. We retained variables in the model if they changed the effect estimates for any of the four UNGD activity metrics by more than 5%. Model fit was assessed by comparing the Akaike's information criterion (AIC) and examining the distribution of model residuals from the fixed effects portion of our models. We did not include models that adjusted for medication use or comorbidities in our main analysis (**Table 6)** because we hypothesized that some of these (e.g., diagnoses of myocardial infarction or hypertension, anti-hypertensive medications) could be both measures of disease severity and potential effect modifiers, and also potentially mediators of the association between UNGD activity and HF hospitalization. Therefore, we evaluated the impact of comorbid conditions and relevant medication use in sensitivity analyses.

3.3.8 Sensitivity analyses

We conducted several sensitivity analyses. First, we evaluated multilevel logistic regression models that included only one random intercept for each subject's place type identifier, whereas our original models included two random intercepts for both place type identifiers and for each subject included in the model (**Table 3.7, Model 1**). Second, to assess the extent to which very young or very old subjects could be influencing our results, we evaluated associations between UNGD activity and hospitalization for heart failure, including only subjects in the age range of 40-80 years (**Table 3.7, Model 2**).

Third, we included a variable for the number of days since a subject's heart failure diagnosis instead of duration of contact with the health care system (**Table 3.7, Model 3**) to evaluate the extent to which our results were affected by differing ways of measuring a subject's duration of care ascertained from the EHR. Fourth, to increase our statistical power to detect an association between UNGD activity and hospitalization for heart failure, we replicated our study sample using a 4:1 instead of a 1:1 matching strategy and evaluated our final models in this 4:1 matched sample using both days since heart failure diagnosis (**Table 3.8, Model 1**) and duration of contact with the health care system (**Table 3.8, Model 2**). Fifth, we evaluated a model that included inverse probability weights (**Table 3.9**) to account for potential selection bias from the 13,183 subjects eligible for either this case-control study or the **Chapter 4** study of laboratory measures (see **Chapter 3.2, Figure 1**). Sixth, to evaluate whether spatial confounding could account for our observed associations, we conducted a negative exposure control analysis (**Figure 3.6**). In this analysis, we assigned UNGD activity metrics in a temporally nonsensical way, such that the UNGD activity metrics could not have caused heart failure hospitalization [229, 230]. We limited our analyses to events from 2008 and 2009, and we assigned UNGD activity to these events from six years after the hospitalization or control encounter date (i.e., 2014 and 2015 UNGD activity). Lastly, we evaluated the impact of medication use (**Table 3.10**) and comorbid diagnoses (**Table 3.11**) on the associations between UNGD activity metrics and hospitalization for heart failure. All analyses were conducted using Stata v13.1 (StataCorp LP 2016. Stata/MP 13.1. College Station, TX.), R (R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria), and ArcGIS (ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute).

3.4 RESULTS

3.4.1 DESCRIPTION OF STUDY SUBJECTS

Of the 12,330 subjects eligible for selection into this study (**Figure 3.1**), 5839 had a first hospitalization for heart failure between 2008-2015, and were thus identified as the cases for this analysis. Because we used a 1:1 control to case frequency matching strategy based on year, sex, and age at event, there were some controls that were used multiple times to meet these criteria (**Table 3.1**). As years progressed in the study period, the mean number of times that subjects without hospitalizations were selected as a control steadily increased (**Table 3.1**). Similarly, the number of subjects eligible for selection into the study increased with year, with the exception of the year 2015, because our 2015 data only included hospitalizations through July 31, 2015. This is also evident in the relatively smaller number of hospitalizations included in 2015 compared to earlier years (**Table 3.1**).

Nearly half (47.5 %) of the eligible heart failure subjects had at least one hospitalization for heart failure during the study period. There were significant differences between subjects with and without hospitalizations for mean duration of contact with the health system and mean distance to major and minor roads, as well as the distribution of subjects by community type, smoking status, and proportion deceased at the end of the study period, according to analysis of variance (ANOVA) (**Table 3.2**). However, the differences in subjects with and without hospitalization for duration of contact; distance to major and minor roads; and smoking status, were not clinically meaningful (**Table 3.2**). Differences between subjects with and without hospitalization by community type, however, appeared to be meaningfully different, with 9.3 % of subjects who were never hospitalized living in census tracts vs. 12.9% who were ever hospitalized for heart failure (**Table 3.2**). This is a major reason for including a random intercept for each subject's community type identifier in primary models of

hospitalization. We also compared the presence of medications and comorbidities by subjects with and without hospitalizations and observed statistically significant differences in antihypertensive medication use, with a higher proportion of usage (44.1%) in subjects who were never hospitalized compared to those who were (39.8%) hospitalized for heart failure (**Table 3.3**). The presence of a number of comorbidities also differed significantly comparing subjects who were hospitalized to those who were not, with a higher proportion of subjects having chronic obstructive pulmonary disease, coronary artery disease, hypertension, valve disorders, type 2 diabetes, and chronic kidney disease, comparing those who were ever hospitalized to those who never were (**Table 3.3**). Lastly, subjects who were hospitalized for heart failure had slightly higher values for the Charlson Index of morbidity, which is driven by the presence of these other comorbidities (**Table 3.3**).

3.4.2 Evaluation of temporal and spatial characteristics of UNGD activity metrics

The proportion of persons in the 4th quartile of the UNGD activity metrics was greatest in the years 2010-2012 for the pad preparation, spud, and stimulation metrics (**Table 3.4**). The production metric was the only metric that did not have a termination date and thus accumulates over time, which is why no subjects were within the 1st quartile of UNGD production activity in the later years of the study (**Table 3.4**). The temporal trends in UNGD activity by phase support the use of our control to case frequency matching strategy, where control encounters were matched to cases based on year of event in addition to sex and age category. Additionally, since the temporal trends in each of the four UNGD metrics were different, estimating the odds of hospitalization in this study separately for each of the four UNGD activity metrics reduces concern about the temporal patterns in each of the metrics.

When we examined the distributions of subjects who were ever hospitalized,

never hospitalized, and those who were selected as a control by UNGD activity metric quartiles (**Table 3.5**), we found that a higher proportion of cases were in the 4th quartile of UNGD activity for the pad preparation metric, the stimulation metric, and the production metric compared to subjects who were never hospitalized and those who were selected as a control. The proportion of subjects in the 4th quartile of the spud (drilling) metric, however, was relatively lower in those who were ever hospitalized compared to those who only served as a control or served as a control and then a case later (**Table 3.5**).

Similar to the temporal patterns in the UNGD activity metrics, spatial patterns varied depending on the metric considered; subjects in the 4th quartile of the spud metric (**Figure 3.3**) were more likely to be located in the central, southwest and northwest regions (**Figure 3.5**), whereas subjects in the 4th quartile of the production metric (**Figure 3.4**) were more concentrated in the northeast, southeast, and central regions. These spatial patterns by UNGD activity metrics motivated us to evaluate how inclusion of a regional variable, which could account for spatial variability at a level beyond each subject's community type or individual county, in our final models of UNGD activity and heart failure hospitalization changed UNGD associations.

3.4.3 Adjusted associations of UNGD activity metrics with heart failure hospitalization

We observed exposure-effect relations, with increasing levels of covariate control, for three of the four UNGD activity metrics with the adjusted odds of heart failure hospitalization (**Table 3.6**). After adjustment for *a priori* covariates (**Table 3.6**, **Model 1**), all four UNGD metrics had at least some significant associations in individual quartiles, but the clearest exposure-effect relations were observed for the stimulation and production metrics. After addition of year of hospitalization or control encounter date

(**Table 3.6, Model 2**), exposure-effect relations for stimulation and production strengthened and one for pad preparation emerged. For the spud metric, a significant protective association emerged in the 4th quartile. After addition of a regional indicator variable (**Figure 3.5**), all associations were attenuated (**Table 3.6, Model 3**), and the 4th quartile protective association for the spud metric was no longer present. When both region and year (**Table 3.6, Model 4**) and then observation time, distance to nearest Geisinger hospital or clinic, and season (**Table 3.6, Model 5**), were added to models, associations with spud were no longer present, while those for pad preparation, stimulation, and production metrics evidenced exposure-effect relations. Additional adjustment for distance to major or minor road did not substantively change either associations or inferences (results not shown).

3.4.5 Sensitivity analyses

Associations from models with only one random intercept for a subject's community identifier (**Table 3.7, Model 1**) or with age restrictions (**Table 3.7, Models 2** and **3**) were inferentially similar to the results from our primary models (**Table 3.6**). We did notice, however, that the adjustment for duration of heart failure (i.e., the date of encounter minus the date of first heart failure diagnosis) in **Table 3.7, Model 3** resulted in smaller effect estimates compared to the use of duration of contact with the health care system (**Table 3.7, Model 2**). However, these differences did not exceed a 10% change in effect estimates and did not impact the inference from these models. Associations from models from 4:1 control to case matching (**Table 3.8**) were substantively similar to those of the primary analysis (**Table 3.6**), with the exception that, in the 4:1 matched analysis, we observed exposure-effect relations for the spud metric (**Table 3.8**). The negative exposure control analysis did not reveal associations between UNGD activity metrics from 2014 and 2015 assigned to events six years before (**Figure**

3.6), although associations for the 4th quartile of UNGD production activity showed increased odds of hospitalization (OR (95% CI): 1.35 (1.00, 1.82).

Additional adjustment for medication use by class (**Table 3.9**) or selected comorbidities (**Table 3.10**) did not substantively change associations or inferences regarding associations between UNGD activity phases, by quartile, and odds of hospitalization. It is worth noting that adjustment for coronary artery disease and for chronic kidney disease (**Table 3.10**) yielded slightly higher odds ratios for hospitalization with each metric of UNGD activity, except for the spud metric, which was consistently null. Adjustment for hypertension diagnoses (**Table 3.10**) also had a minor effect on the odds ratios, which slightly reduced the odds ratios for hospitalization across all metrics.

3.5 DISCUSSION

The goal of this analysis was to evaluate the adjusted associations between metrics of UNGD activity and hospitalization among subjects with heart failure in a large region of UNGD activity over approximately 10 years. To our knowledge, this question had never before been evaluated. The study was motivated by strong biologic rationale and *a priori* hypotheses regarding how the environmental impacts of UNGD [2, 17, 60] could affect cardiovascular health. Our findings support candidate exposure pathways of air pollution and stress, and they suggest that individuals living with heart failure are more likely to be hospitalized when exposed to greater UNGD activity. To our knowledge, this is the first study of UNGD in relation to a health outcome that primarily affects older persons, who likely would be more susceptible to the exposures associated with UNGD. The findings are thus an interesting and important contribution to the expanding literature of UNGD and health.

We observed exposure-effect relations with quartile of UNGD activity and odds of hospitalization for heart failure. These effects were robust to increasing spatiotemporal

covariate adjustment and several sensitivity analyses. The strongest and most robust associations (OR [95% CI]) were observed in the 4th (vs. 1st) quartile of the pad preparation, stimulation, and production metrics, as 1.70 (95% CI: 1.35, 2.13), 1.80 (95% CI: 1.35, 2.40), and 1.62 (95% CI: 1.07, 2.45), respectively (**Table 3.6, Model 5**). Because our UNGD activity metrics had strong temporal trends (**Table 3.4**) and spatial patterns (**Figures 3.3** and **3.4**), we believe that the adjustment for year and region is necessary and sufficient given that quartiles of UNGD activity were well represented across regions. The effect of adjusting for these variables is especially evident in the differing associations between **Models 1-4** in **Table 3.6** which we believe reflect the underlying spatiotemporal trends in the data. In **Models 1-5**, we still observed consistent exposure-effect relations with quartile of UNGD activity for the pad preparation, stimulation, and production metrics, across all models. We consistently found null associations with the spud metric, except for in our 4:1 control to case matched sensitivity analysis.

To assess whether unmeasured spatial confounding could account for our associations, we completed a negative exposure control analysis, assigning UNGD activity metrics from 2014 and 2015 to hospitalizations and matched control dates in 2008 and 2009 (**Figure 3.6**). If there was time-invariant spatial confounding, that is, features of people in places or places themselves that differed geographically, we would expect such an analysis to show associations of UNGD activity with case status even though the causal temporality requirement would be violated. In the negative exposure control analysis, UNGD activity assigned in this way was not associated with earlier hospitalizations. We did not perform negative outcome control analyses (i.e., evaluating UNGD activity in relation to a biologically implausible outcome) because these have been conducted in this region using the same UNGD activity metrics and Geisinger data as used in this study, and all have found null associations, for diarrheal illness [15], skin

and soft tissue infections [13], and cold/flu, ear pain, or bad breath symptoms [17].

We had initially suspected that medication classes and comorbidities would modify associations between UNGD activity and hospitalization for heart failure. As effect modifiers, medications and co-occurring conditions (such as those listed in **Table 3.3**) could theoretically impact the association between UNGD activity and hospitalization. However, models that adjusted for each of these medication classes (**Table 3.9**) and comorbid conditions (**Table 3.10**) individually did not substantially change inference of the exposure-effect relations between UNGD activity and hospitalization for heart failure, and we did not observe any effect modification by medication use in models that included UNGD cross products. Models that adjusted for coronary artery disease and for chronic kidney disease resulted in slightly higher odds ratios for hospitalization by quartile of UNGD activity (**Table 3.11**), but this is probably because both of these conditions have an independent association with hospitalization. Across all models, including those from the several sensitivity analyses (**Tables 3.7-3.10, Figure 3.6**), we consistently saw exposure-effect relations for the pad preparation, stimulation, and production metrics and hospitalization.

The body of epidemiologic evidence demonstrating that UNGD activity is associated with adverse health outcomes is growing. The findings from this study are consistent with the majority of epidemiology studies on UNGD in Pennsylvania demonstrating biologically-plausible associations with health outcomes. These include nasal and sinus, migraine headache, and fatigue symptoms (Tustin, Hirsch et al. 2016), asthma exacerbations [15], depressive symptoms [18], respiratory and skin symptoms [63], and adverse birth outcomes [9, 10, 13, 17]). Similarly, recent studies have found associations between UNGD and quality of life [231], preterm birth [11], and inflammatory biomarkers relevant to cardiovascular disease [66] in Ohio, Texas, and Colorado, respectively. The findings of this study add to this growing body of evidence

and provides further evidence that UNGD is associated with a range of health impacts, which now includes hospitalizations for heart failure among heart failure subjects.

3.5.1 Strengths and limitations

This study had the advantage of utilizing EHR data from a large, representative subject population in Pennsylvania living with varying intensity of UNGD activity over an eight-year study period. First, our inverse-distance weighted UNGD activity metrics were developed directly from data from Pennsylvania's Department of Environmental Protection and are more reflective of intensity of UNGD activity than crude measures (i.e., distance buffers), which is an advantage for use in epidemiologic studies [190]. Second, we had a large sample size of 9054 subjects and 5839 hospitalizations. Third, because we had access to EHR information, we were able to understand each subject's contact with the medical system, visit history, comorbidities and medications. Because the occurrence of multiple comorbidities is common in subjects with heart failure (**Table 3.3**), we were able to assess the extent to which the associations between UNGD activity and hospitalization were affected by additional adjustment for these medications and comorbidities (**Tables 3.9 & 3.10**). Fourth, we employed several other sensitivity analyses (e.g., restricting by age, utilizing a 4:1 matching strategy, inverse probability of selection weighting, using only one vs. two random intercepts in our models, a negative exposure control analysis) to assess the robustness of our primary models (**Table 3.6**), and none of these sensitivity analyses suggested that the association between UNGD activity and hospitalization was a result of unmeasured confounding.

The study also had some limitations. Importantly, we did not have information on dietary intake, physical activity, and current or past occupations. Although there is some information on alcohol use available from the EHR, we did not consider alcohol use in this analysis because there is evidence that US adults, particularly older adults, do not

frequently discuss alcohol use with their provider [232]. This is evident in the EHR social history file's data. Unlike information on tobacco smoking, we had limited information that our subjects used alcohol, with only 3966 of the 9054 subjects confirming alcohol use at any given time point. Of those with available alcohol use information, it was often not reported at the time of the hospitalization or control encounter. We also did not have information on occupation or occupational status, however we do not have any information to suggest that current or past occupation should be highly correlated with our UNGD activity metrics, and do not believe that most of these subjects with heart failure were likely to have been employed currently or in the recent past. Lastly, we did not have data for individuals who do not seek care from the health system, so there is potential for unmeasured differences in persons who sought care from the health system vs. those who did not. This is not a serious concern given the severity of heart failure and the high likelihood that care would be required.

3.6 Conclusions and future directions

To the best of our knowledge, this is the first large scale epidemiologic study to evaluate associations of UNGD activity with heart failure hospitalization among subjects with heart failure. We observed significantly increased odds of hospitalization among heart failure subjects in relation to increasing UNGD activity for several phases, including pad preparation, stimulation, and production. These associations are plausible given the documented environmental impacts of UNGD (e.g., air pollution [233], water contamination [45], noise [3], traffic [6], and community impacts [54, 234]. The findings of this study support candidate exposure pathways involving air pollution and stress, as these are the most biologically plausible pathways to explain associations with heart failure hospitalization. Understanding how people living with heart failure are susceptible to environmental exposures, especially those associated with UNGD, is especially

important given the growing prevalence of heart failure [89]. Future studies should better characterize mechanistic pathways underlying these observed associations (i.e., differentiate between air pollution, stress, or other potential exposure pathways). In the next chapter, I evaluate associations between UNGD activity and laboratory measures of B-type natriuretic peptide (BNP) in blood, a clinical biomarker frequently used in heart failure diagnosis and management, and one that I hypothesize could be a marker of an early biologic effect associated with UNGD activity exposure.

3.7 REFERENCES

1. Mozaffarian, D., et al., *Heart Disease and Stroke Statistics—2016 Update: A Report From the American Heart Association.* Circulation, 2016. **133**(4): p. e38-e360.
2. CDC;. *Heart Failure Fact Sheet*. 2013 [cited 2015 October]; Available from: http://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_failure.htm.
3. Savarese, G. and L.H. Lund, *Global Public Health Burden of Heart Failure.* Cardiac Failure Review, 2017. **3**(1): p. 7-11.
4. Heidenreich, P.A., et al., *Forecasting the Impact of Heart Failure in the United States: A Policy Statement From the American Heart Association.* Circulation: Heart Failure, 2013.
5. Bui, A.L., T.B. Horwich, and G.C. Fonarow, *Epidemiology and risk profile of heart failure.* Nature reviews. Cardiology, 2011. **8**(1): p. 30-41.
6. Salamanca-Bautista, P., et al., *Precipitating factors of heart failure admission: Differences related to age and left ventricular ejection fraction.* (1874-1754 (Electronic)).
7. Sin, D.D. and S.F.P. Man, *Chronic Obstructive Pulmonary Disease as a Risk Factor for Cardiovascular Morbidity and Mortality.* Proceedings of the American Thoracic Society, 2005. **2**(1): p. 8-11.
8. Friedmann, E., et al., *Relationship of depression, anxiety, and social isolation to chronic heart failure outpatient mortality.* American Heart Journal, 2006. **152**(5): p. 940.e1-940.e8.
9. Ogilvie, R.P., et al., *Psychosocial Factors and Risk of Incident Heart Failure: The Multi-Ethnic Study of Atherosclerosis.* Circ Heart Fail, 2016. **9**(1): p. e002243.
10. Yohannes, A.M., et al., *Depression and anxiety in chronic heart failure and chronic obstructive pulmonary disease: prevalence, relevance, clinical implications and management principles.* International Journal of Geriatric Psychiatry, 2010. **25**(12): p. 1209-1221.
11. Benjamin, E.J., et al., *Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association.* Circulation, 2017. **135**(10): p. e146-e603.
12. Heidenreich, P.A., et al., *Forecasting the Future of Cardiovascular Disease in the United States.* Circulation, 2011. **123**(8): p. 933.
13. Jackson Sandra, L., et al., *National Burden of Heart Failure Events in the United States, 2006 to 2014.* Circulation: Heart Failure, 2018. **11**(12): p. e004873.
14. von Lueder, T.G. and S. Agewall, *The burden of heart failure in the general population: a clearer and more concerning picture.* Journal of thoracic disease, 2018. **10**(Suppl 17): p. S1934-S1937.
15. Zannad, F., *Rising incidence of heart failure demands action.* The Lancet, 2018. **391**(10120): p. 518-519.
16. Harjola, V.-P., et al., *Organ dysfunction, injury and failure in acute heart failure: from pathophysiology to diagnosis and management. A review on behalf of the Acute Heart Failure Committee of the Heart Failure Association (HFA) of the European Society of Cardiology (ESC).* European journal of heart failure, 2017. **19**(7): p. 821-836.
17. Ponikowski, P., et al., *2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure.* European Journal of Heart Failure, 2016. **18**(8): p. 891-975.
18. Bhatnagar, A., *Environmental Cardiology.* Circulation Research, 2006. **99**(7): p. 692.

19. Shah, A.S.V., et al., *Global association of air pollution and heart failure: a systematic review and meta-analysis.* The Lancet. **382**(9897): p. 1039-1048.
20. Sorensen, M., et al., *Long-Term Exposure to Road Traffic Noise and Nitrogen Dioxide and Risk of Heart Failure: A Cohort Study.* (1552-9924 (Electronic)).
21. Levin, R.K., et al., *Increased hospitalizations for decompensated heart failure and acute myocardial infarction during mild winters: A seven-year experience in the public health system of the largest city in Latin America.* (1932-6203 (Electronic)).
22. Wellenius, G.A., et al., *Particulate Air Pollution and the Rate of Hospitalization for Congestive Heart Failure among Medicare Beneficiaries in Pittsburgh, Pennsylvania.* American Journal of Epidemiology, 2005. **161**(11): p. 1030-1036.
23. Brook, R.D., et al., *Air Pollution and Cardiovascular Disease.* Circulation, 2004. **109**(21): p. 2655.
24. Münzel, T., et al., *Environmental stressors and cardio-metabolic disease: part I–epidemiologic evidence supporting a role for noise and air pollution and effects of mitigation strategies.* European Heart Journal, 2017. **38**(8): p. 550-556.
25. Schmidt, F., et al., *Nighttime aircraft noise impairs endothelial function and increases blood pressure in patients with or at high risk for coronary artery disease.* Clinical Research in Cardiology, 2015. **104**(1): p. 23-30.
26. Hagström, E., et al., *Psychosocial stress and major cardiovascular events in patients with stable coronary heart disease.* Journal of Internal Medicine, 2018. **283**(1): p. 83-92.
27. Pennsylvania Department of Environmental Protection, *PA DEP Oil & Gas Reporting Website* 2016.
28. McKenzie, L.M., et al., *Human health risk assessment of air emissions from development of unconventional natural gas resources.* Sci Total Environ, 2012. **424**: p. 79-87.
29. Maryland Department of Environment and Maryland Department of Natural Resources, *Assessment of Risks from Unconventional Gas Well Development in the Marcellus Shale of Western Maryland*. 2015.
30. Swarthout, R.F., et al., *Impact of Marcellus Shale Natural Gas Development in Southwest Pennsylvania on Volatile Organic Compound Emissions and Regional Air Quality.* Environmental Science & Technology, 2015. **49**(5): p. 3175-3184.
31. Johnson, D., et al., *Estimated Emissions from the Prime-Movers of Unconventional Natural Gas Well Development Using Recently Collected In-Use Data in the United States.* Environmental Science & Technology, 2018.
32. Stacy, S., et al., *Perinatal outcomes and unconventional natural gas operations in Southwest Pennsylvania.* PLoS One, 2015. **10**(6): p. e0126425.
33. Currie, J., M. Greenstone, and K. Meckel, *Hydraulic fracturing and infant health: New evidence from Pennsylvania.* Science Advances, 2017. **3**(12): p. e1603021.
34. Whitworth, K.W., A.K. Marshall, and E. Symanski, *Maternal residential proximity to unconventional gas development and perinatal outcomes among a diverse urban population in Texas.* PloS one, 2017. **12**(7): p. e0180966-e0180966.
35. Walker Whitworth, K., A. Kaye Marshall, and E. Symanski, *Drilling and Production Activity Related to Unconventional Gas Development and Severity of Preterm Birth.* 2018(1552-9924 (Electronic)).
36. Casey, J.A., et al., *Unconventional Natural Gas Development and Birth Outcomes in Pennsylvania, USA.* Epidemiology, 2015.
37. Janitz, A.E., et al., *The association between natural gas well activity and specific congenital anomalies in Oklahoma, 1997–2009.* Environment International, 2019. **122**: p. 381-388.

38. Rasmussen, S.G., et al., *Association Between Unconventional Natural Gas Development in the Marcellus Shale and Asthma Exacerbations. .* JAMA Internal Medicine, 2016(2168-6114).
39. Willis, M.D., et al., *Unconventional natural gas development and pediatric asthma hospitalizations in Pennsylvania.* (1096-0953 (Electronic)).
40. Casey, J.A., et al., *Associations of unconventional natural gas development with depression symptoms and disordered sleep in Pennsylvania.* Scientific Reports, 2018. **8**(1): p. 11375.
41. Adgate, J.L., B.D. Goldstein, and L.M. McKenzie, *Potential Public Health Hazards, Exposures and Health Effects from Unconventional Natural Gas Development.* Environmental Science & Technology, 2014. **48**(15): p. 8307-8320.
42. McKenzie, L.M., et al., *Relationships between indicators of cardiovascular disease and intensity of oil and natural gas activity in Northeastern Colorado.* Environmental Research, 2019. **170**: p. 56-64.
43. Casey, J.A., et al., *Using Electronic Health Records for Population Health Research: A Review of Methods and Applications.* Annual Review of Public Health, 2016. **37**(1): p. 61-81.
44. Wang, M.-H., et al., *A Simulation Study of Control Sampling Methods for Nested Case-Control Studies of Genetic and Molecular Biomarkers and Prostate Cancer Progression.* Cancer Epidemiology Biomarkers & Prevention, 2009. **18**(3): p. 706.
45. Quan, H., et al., *Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10.* Med Care, 2005. **43**(11): p. 1130-9.
46. Casey, J.A., et al., *Measures of SES for Electronic Health Record-based Research.* American Journal of Preventive Medicine, 2018. **54**(3): p. 430-439.
47. Nau, C., et al., *Community socioeconomic deprivation and obesity trajectories in children using electronic health records.* Obesity (Silver Spring), 2015. **23**(1): p. 207-12.
48. Greene, S.J., et al., *Medical Therapy for Heart Failure With Reduced Ejection Fraction.* Journal of the American College of Cardiology, 2018. **72**(4): p. 351.
49. Hood, S.R., et al., *Association Between Medication Adherence and the Outcomes of Heart Failure.* Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 2018. **38**(5): p. 539-545.
50. Tschöpe, C., et al., *Heart failure with preserved ejection fraction: current management and future strategies.* Clinical Research in Cardiology, 2018. **107**(1): p. 1-19.
51. Linting, M., R.C.A. Rippe, and A. van der Voort, *Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data AU - van Ginkel, Joost R.* Journal of Personality Assessment, 2019: p. 1-12.
52. Koehler, K., et al., *Exposure Assessment Using Secondary Data Sources in Unconventional Natural Gas Development and Health Studies.* Environmental Science & Technology, 2018. **52**(10): p. 6061-6069.
53. Liu, A.Y., et al., *The contextual influence of coal abandoned mine lands in communities and type 2 diabetes in Pennsylvania.* Health & Place, 2013. **22**: p. 115-122.
54. Schwartz, B.S., et al., *Body Mass Index and the Built and Social Environments in Children and Adolescents Using Electronic Health Records.* American Journal of Preventive Medicine, 2011. **41**(4): p. e17-e28.
55. *American Community Survey*. 2010, United States Census Bureau.

56. Tustin, A.W., et al., *Associations between Unconventional Natural Gas Development and Nasal and Sinus, Migraine Headache, and Fatigue Symptoms in Pennsylvania.* Environ Health Perspect, 2016.
57. Casey, J., A., et al., *Predictors of Indoor Radon Concentrations in Pennsylvania, 1989–2013.* Environmental Health Perspectives, 2015. **123**(11): p. 1130-1137.
58. Chang, J.H., et al., *Association of time-serial changes in ambient particulate matters (PMs) with respiratory emergency cases in Taipei's Wenshan District.* 2017(1932-6203 (Electronic)).
59. Shah, A.S., et al., *Short term exposure to air pollution and stroke: systematic review and meta-analysis.* (1756-1833 (Electronic)).
60. Lo, A.X., et al., *A National Study of U.S. Emergency Departments: Racial Disparities in Hospitalizations for Heart Failure.* (1873-2607 (Electronic)).
61. Weisskopf, M.G., R. Tchetgen Tchetgen Ej Fau - Raz, and R. Raz, *Commentary: On the Use of Imperfect Negative Control Exposures in Epidemiologic Studies.* (1531-5487 (Electronic)).
62. Lipsitch, M., E. Tchetgen Tchetgen, and T. Cohen, *Negative controls: a tool for detecting confounding and bias in observational studies.* Epidemiology (Cambridge, Mass.), 2010. **21**(3): p. 383-388.
63. Union of Concerned Scientists, *Toward an Evidence-Based Fracking Debate: Science, Democracy, and Community Right to Know in Unconventional Oil and Gas Development*. 2013.
64. Vengosh, A., et al., *A Critical Review of the Risks to Water Resources from Unconventional Shale Gas Development and Hydraulic Fracturing in the United States.* Environmental Science & Technology, 2014. **48**(15): p. 8334-8348.
65. Rabinowitz, P.M., et al., *Proximity to natural gas wells and reported health status: results of a household survey in Washington County, Pennsylvania.* Environ Health Perspect, 2015. **123**(1): p. 21-6.
66. Fisher, M.P., et al., *Psychosocial implications of unconventional natural gas development: Quality of life in Ohio's Guernsey and Noble Counties.* Journal of Environmental Psychology, 2018. **55**: p. 90-98.
67. McKnight-Eily, L.R., et al., *Vital Signs: Communication Between Health Professionals and Their Patients About Alcohol Use.* Morbidity and Mortality Weekly Report, 2014. **63**(1): p. 16-22.
68. Srebotnjak, T. and M. Rotkin-Ellman, *Fracking Fumes: Air pollution from hydraulic fracturing threatens public health and communities*. 2014, Natural Resources Defense Council;.
69. Osborn, S.G., et al., *Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing.* Proceedings of the National Academy of Sciences, 2011. **108**(20): p. 8172-8176.
70. Hays, J. and S.B.C. Shonkoff, *Toward an Understanding of the Environmental and Public Health Impacts of Unconventional Natural Gas Development: A Categorical Assessment of the Peer-Reviewed Scientific Literature, 2009-2015.* PLoS ONE, 2016. **11**(4): p. e0154164.
71. Graham, J., et al., *Increased traffic accident rates associated with shale gas drilling in Pennsylvania.* Accident Analysis & Prevention, 2015. **74**: p. 203-209.
72. Perry, S.L., *Development, Land Use, and Collective Trauma: The Marcellus Shale Gas Boom in Rural Pennsylvania.* Culture, Agriculture, Food and Environment, 2012. **34**(1): p. 81-92.
73. Muehlenbachs, L., E. Spiller, and C. Timmins, *The Housing Market Impacts of Shale Gas Development.* The American Economic Review, 2015. **105**(12): p. 3633-3659.

Figure 3.1. Flowchart detailing study selection from electronic health records

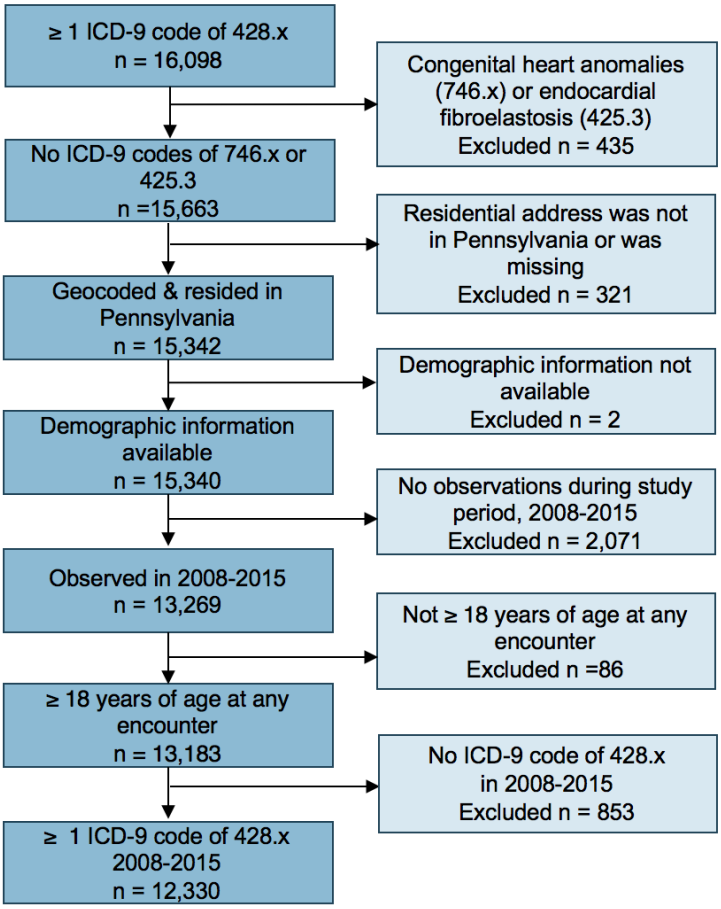


Figure 3.2. Timeline of UNGD activity assignment in relation to events (hospitalizations or control encounters)

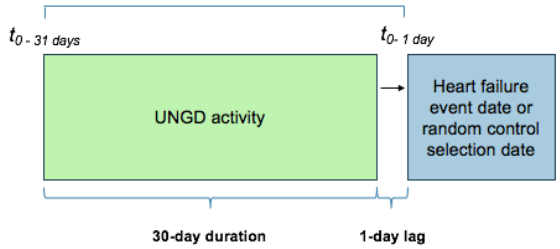


Table 3.1. Description of cases and controls eligible for selection, and selected, by year, for the primary analysis of 1:1 case to control matching

	Under observation (n)				Selected as Control (n)		
Year	Not new* this year	New this year	Become case (n)	Available for control selection (n)**	Not a case later	Case later	Mean (SD) control selection number †
2008	0	3,966	670	3296	188	482	1 (0)
2009	3,452	1,049	606	3895	187	419	1.10 (0.31)
2010	3,572	932	582	3922	204	378	1.28 (0.50)
2011	3,719	890	579	4030	248	331	1.76 (0.76)
2012	3,975	1,152	765	4362	373	779	1.81 (0.84)
2013	4,353	1,415	964	4804	577	387	1.85 (0.90)
2014	6,768	1,604	1,143	5164	921	222	1.90 (0.93)
2015	4,519	714	530	4703	530	0	1.97 (1.05)

* Subjects who became eligible for control selection in the corresponding year
** Heart failure subjects who did not have a heart failure hospitalization
† Cumulative mean number of times controls who were selected had been a control

Table 3.2. Selected subject characteristics by case and control status

	Never a case	Ever a case	
Total subjects n = 9054	**n = 3215**	**n = 5839**	**p-value***
Sex, n (%)			
Male	1659 (51.6%)	3074 (52.7%)	
Female	1556 (48.4%)	2765 (47.4%)	0.3
Age at hospitalization or at control selection date, years, mean (SD)	71.0 (12.6)	71.1 (12.7)	0.8
Age category at first event, n (%) years			
> 18-30	20 (0.6)	41 (0.7)	
> 30-40	37 (1.2)	72 (1.2)	
> 40-50	152 (4.7)	245 (4.2)	
> 50-60	420 (13.1)	737 (12.6)	
> 60-70	727 (22.6)	1331 (22.8)	
> 70-80	916 (28.5)	1734 (29.7)	
> 80-90	903 (28.1)	1608 (27.5)	
> 90-100	40 (1.2)	71 (1.2)	0.9
Race/ethnicity, n (%)			
White	3127 (97.3%)	5688 (97.4%)	
Black	46 (1.4%)	74 (1.3%)	
Hispanic	28 (0.9%)	60 (1.0%)	
Other	10 (0.3%)	17 (0.3%)	
Missing	4 (0.1%)	0 (0%)	0.08
Smoking status at event, n (%)			
Current	358 (11.1)	760 (13.0)	
Former	1539 (47.9)	2690 (46.1)	
Never	1318 (41.0)	2389 (40.9)	0.03

	Never a case	Ever a case	
Total subjects n = 9054	**n = 3215**	**n = 5839**	**p-value***
Community type, n (%)			
Borough	991 (30.8%)	1834 (31.4%)	
Township	1926 (59.9%)	3248 (55.6%)	
Census tract (city)	298 (9.3%)	757 (12.9%)	< 0.001
CSD,** SD units, quartiles			
1 (-7.5, -2.6)	615 (19.1%)	1081 (18.5%)	
2 (-2.6, -0.5)	823 (25.6%)	1419 (24.3%)	
3 (-0.5, 2.3)	904 (28.1%)	1775 (30.4%)	
4 (2.3, 22.6)	873 (27.2%)	1564 (26.8%)	0.13
Patient status at end of study, n (%)			
Alive	2514 (78.2%)	3487 (59.7%)	
Deceased	701 (21.8%)	2352 (40.3%)	< 0.001
Distance to major road † (meters), mean (SD)	2908 (4,160)	2703 (4,282)	0.03
Distance to minor road † (meters), mean (SD)	1784 (2599)	1431 (2136)	< 0.001
Receipt of Medical Assistance‡ n (%)	368 (11.5)	681 (11.7)	0.8
Body mass index (BMI) at event, kg/m^2, mean (SD)	31.9 (7.7)	31.5 (8.8)	0.5
Duration of contact with health system, days§, mean (SD)	3995 (1395)	3683 (1530)	< 0.001

*p-value obtained from chi^2 tests comparing selected variable in cases and controls for categorical or binary variables; ANOVA F-test for continuous variables

**Community socioeconomic deprivation (CSD) was calculated based on US Census indicators; further information is detailed in the text

†Major and minor roads were identified from the Pennsylvania Department of Transportation (DOT) databases; distance from subject's residential address to these roads was calculated in meters using the Generate Near Table tool function in ArcGIS 10.4.

‡ Medical Assistance, a surrogate for family socioeconomic status, was calculated based on health insurance status at the time of encounters as previously reported

§ Days from first to most recent (i.e., case event or control selection date) time a subject was observed in the EHR

Table 3.3. Selected diagnoses and medication use at time of hospitalization or control selection day, by heart failure hospitalization status

	Never a case	Ever a case	
Total subjects n = 9054	**n = 3215**	**n = 5839**	**p-value***
Medication use, by class, n (%)**	1418 (44.1)	2324 (39.8)	< 0.001
Antihypertensive	1503 (46.8)	2615 (44.8)	0.07
Antihyperlipidemic	625 (19.4)	1049 (18.0)	0.08
Anticoagulant			
Chronic obstructive pulmonary disease (COPD), n (%)	421 (13.1)	1106 (18.9)	< 0.001
Coronary artery disease, n (%)	492 (15.3)	1220 (20.9)	< 0.001
Hypertension, n (%)	1801 (56.0)	3962 (67.9)	< 0.001
Myocardial infarction, n (%)	212 (6.59)	433 (7.42)	0.2
Valve disorder, n (%)	496 (15.4)	1127 (19.3)	< 0.001
Diabetes, n (%)	989 (30.8)	2356 (40.3)	< 0.001
Chronic kidney disease, n (%)	610 (26.4)	1519 (29.6)	0.001
Charlson Index of morbidity †, mean (SD)	8.26 (3.18)	8.76 (3.36)	0.0001

*p-value obtained from chi^2 tests comparing selected variable in cases and controls for categorical or binary variables; t-test for continuous variables

** Relevant medication classes were identified based on the dates of physician orders

† A composite measure of overall morbidity; definition described in text

Table 3.4. Frequencies of subjects by UNGD activity metric quartile and year

Pad preparation metric					
Year	**Quartile 1 n (%)**	**Quartile 2 n (%)**	**Quartile 3 n (%)**	**Quartile 4 n (%)**	**Total subjects (n)**
2008	589 (44.0)	560 (41.8)	147 (11.0)	44 (3.3)	1,340
2009	221 (18.2)	314 (25.9)	371 (30.6)	306 (25.3)	1,212
2010	11 (1.0)	84 (7.2)	294 (25.3)	775 (66.6)	1,164
2011	5 (0.4)	23 (2.0)	355 (30.7)	775 (66.9)	1,158
2012	33 (2.2)	347 (22.7)	573 (37.5)	577 (37.7)	1,530
2013	411 (21.3)	664 (34.4)	585 (30.3)	268 (13.9)	1,928
2014	789 (34.5)	790 (34.6)	550 (24.1)	157 (6.9)	2,286
2015	861 (81.2)	137 (12.9)	45 (4.2)	17 (1.6)	1,060
Spud metric					
Year	**Quartile 1 n (%)**	**Quartile 2 n (%)**	**Quartile 3 n (%)**	**Quartile 4 n (%)**	**Total subjects (n)**
2008	1,340 (100)	0	0	0	1,340
2009	904 (75.6)	247 (20.4)	44 (3.6)	17 (1.4)	1,212
2010	74 (6.4)	288 (24.7)	419 (36.0)	383 (32.9)	1,164
2011	0	131 (11.3)	244 (21.1)	783 (67.6)	1,158
2012	135 (8.8)	441 (28.8)	486 (31.8)	468 (30.6)	1,530
2013	104 (5.4)	835 (43.3)	619 (32.1)	370 (19.2)	1,928
2014	37 (1.6)	481 (21.0)	929 (40.6)	839 (36.7)	2,286
2015	326 (30.8)	496 (46.8)	179 (16.9)	59 (5.6)	1,060
Stimulation metric					
Year	**Quartile 1 n (%)**	**Quartile 2 n (%)**	**Quartile 3 n (%)**	**Quartile 4 n (%)**	**Total subjects (n)**
2008	1,298 (96.8)	31 (2.3)	5 (0.4)	6 (0.5)	1,340
2009	1,073 (88.5)	105 (8.7)	16 (1.3)	18 (1.5)	1,212
2010	250 (21.5)	472 (40.6)	321 (27.6)	121 (10.4)	1,164
2011	9 (0.8)	298 (25.7)	468 (40.4)	383 (33.1)	1,158
2012	9 (0.6)	230 (15.0)	537 (35.1)	754 (49.3)	1,530
2013	21 (1.1)	568 (29.5)	636 (33.0)	703 (36.5)	1,928
2014	39 (1.7)	728 (31.9)	733 (32.1)	786 (34.4)	2,286
2015	221 (20.9)	487 (45.9)	204 (19.3)	148 (14.0)	1,060
Production metric					
Year	**Quartile 1 n (%)**	**Quartile 2 n (%)**	**Quartile 3 n (%)**	**Quartile 4 n (%)**	**Total subjects (n)**
2008	1,340 (0)	0	0	0	1,340
2009	1,138 (93.9)	68 (5.6)	4 (0.3)	2 (0.2)	1,212
2010	433 (37.2)	703 (60.4)	17 (1.5)	11 (1.0)	1,164
2011	9 (0.8)	966 (83.4)	140 (12.1)	43 (3.7)	1,158

2012	0	608 (39.7)	698 (45.6)	224 (14.6)	1,530
2013	0	339 (17.6)	787 (40.8)	802 (41.6)	1,928
2014	0	167 (7.3)	851 (37.2)	1,268 (55.5)	2,286
2015	0	68 (6.4)	423 (39.9)	569 (53.7)	1,060

Table 3.5. UNGD activity metrics by quartile and phase, by heart failure hospitalization status at date of hospitalization or control selection date

	Ever a case*	Never a case**	Ever a control†
Total subjects, n = 9,054	**n = 5839 (%)**	**n = 3215 (%)**	**n = 5839 (%)**
Pad preparation, $1/m^2$			
Q1 (4.0×10^{-9}, 3.3×10^{-8})	1,309 (22.4%)	900 (28.0%)	1,621 (27.8%)
Q2 (3.3×10^{-8}, 6.2×10^{-8})	1,457 (25.0%)	829 (25.8%)	1,490 (25.5%)
Q3 (6.2×10^{-8}, 1.2×10^{-7})	1,533 (26.3%)	747 (23.2%)	1,348 (23.1%)
Q4 (1.2×10^{-7}, 2.5×10^{-5})	1,540 (26.4%)	739 (23.0%)	1,380 (23.6%)
Spud (drilling), $1/m^2$			
Q1 (1.9×10^{-10}, 2.6×10^{-9})	1,750 (30.0%)	833 (25.9%)	1,416 (24.3%)
Q2 (2.6×10^{-9}, 3.8×10^{-9})	1,513 (25.9%)	677 (21.1%)	1,343 (23.0%)
Q3 (3.8×10^{-9}, 4.7×10^{-8})	1,474 (25.2%)	702 (21.8%)	1,353 (23.2%)
Q4 (4.7×10^{-9}, 1.7×10^{-8})	1,102 (18.9%)	1,003 (31.2%)	1,727 (29.6%)
Stimulation, m/m^2			
Q1 (2.5×10^{-6}, 2.7×10^{-4})	1,679 (28.8%)	907 (28.2%)	1,504 (25.8%)
Q2 (2.7×10^{-4}, 5.6 X 10-4)	1,145 (19.6%)	952 (29.6%)	1,716 (29.4%)
Q3 (5.6×10^{-4}, 9.7×10^{-4})	1,485 (25.4%)	678 (21.1%)	1,323 (22.7%)
Q4 (9.7×10^{-4}, 0.2)	1,530 (26.2%)	679 (21.1%)	1,296 (22.2%)
Production, m^3/m^2			
Q1 (2.2×10^{-6}, 0.002)	1,749 (30.0%)	880 (27.4%)	1,461 (25.0%)
Q2 (0.002, 0.02)	1,255 (21.5%)	923 (28.7%)	1,654 (28.3%)
Q3 (0.02, 0.03)	1,343 (23.0%)	795 (24.7%)	1,477 (25.3%)
Q4 (0.03, 16.5)	1,492 (25.6%)	617 (19.2%)	1,247 (21.4%)

* Distribution of UNGD activity metrics assigned at date of hospitalization

**Distribution of UNGD activity metrics at first randomly selected control encounter date

† Distribution of UNGD activity metrics for all control encounter dates

Table 3.6. Associations of UNGD activity metrics, by phase, with hospitalization for heart failure, from models with increasing covariate adjustment

	Model 1*	Model 2**	Model 3§	Model 4†	Model 5‡
UNGD Metric	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**
Pad preparation, $1/m^2$					
Q1 (4.0×10^{-9}, 3.3×10^{-8})	(ref)	(ref)	(ref)	(ref)	(ref)
Q2 (3.3×10^{-8}, 6.2×10^{-8})	1.07 (0.92, 1.25)	1.23 (1.04, 1.46)	1.03 (0.88, 1.19)	1.16 (0.98, 1.37)	1.19 (1.01, 1.40)
Q3 (6.2×10^{-8}, 1.2×10^{-7})	1.39 (1.19, 1.62)	1.75 (1.44, 2.13)	1.34 (1.15, 1.57)	1.65 (1.36, 2.00)	1.63 (1.35, 1.97)
Q4 (1.2×10^{-7}, 2.5×10^{-5})	1.29 (1.10, 1.52)	1.84 (1.45, 2.33)	1.24 (1.06, 1.46)	1.70 (1.34, 2.15)	1.70 (1.35, 2.13)
Spud (drilling), $1/m^2$					
Q1 (1.9×10^{-10}, 2.6×10^{-9})	(ref)	(ref)	(ref)	(ref)	(ref)
Q2 (2.6×10^{-9}, 3.8×10^{-9})	1.26 (1.08, 1.47)	0.94 (0.76, 1.16)	1.23 (1.06, 1.43)	1.01 (0.82, 1.25)	1.01 (0.82, 1.25)
Q3 (3.8×10^{-9}, 4.7×10^{-8})	1.26 (1.09, 1.47)	0.90 (0.71, 1.13)	1.27 (1.09, 1.48)	1.04 (0.82, 1.31)	1.07 (0.85, 1.35)
Q4 (4.7×10^{-9}, 1.7×10^{-8})	0.96 (0.82, 1.13)	0.64 (0.49, 0.83)	1.16 (0.99, 1.37)	0.93 (0.71, 1.21)	0.97 (0.75, 1.27)
Stimulation, m/m^2					
Q1 (2.5×10^{-6}, 2.7×10^{-4})	(ref)	(ref)	(ref)	(ref)	(ref)
Q2 (2.7×10^{-4}, 5.6 X 10-4)	0.96 (0.82, 1.12)	1.03 (0.80, 1.31)	0.99 (0.85, 1.16)	1.00 (0.78, 1.28)	1.03 (0.81, 1.31)
Q3 (5.6×10^{-4}, 9.7×10^{-4})	1.45 (1.24, 1.70)	1.67 (1.27, 2.20)	1.41 (1.20, 1.65)	1.50 (1.14, 1.98)	1.56 (1.19, 2.04)
Q4 (9.7×10^{-4}, 0.2)	1.65 (1.39, 1.96)	1.98 (1.47, 2.67)	1.62 (1.37, 1.92)	1.78 (1.32, 2.40)	1.80 (1.35, 2.40)
Production, m^3/m^2					
Q1 (2.2×10^{-6}, 0.002)	(ref)	(ref)	(ref)	(ref)	(ref)
Q2 (0.002, 0.02)	1.01 (0.86, 1.18)	0.89 (0.65, 1.21)	1.07 (0.92, 1.26)	0.83 (0.61, 1.14)	0.87 (0.64, 1.19)
Q3 (0.02, 0.03)	1.19 (1.01, 1.40)	1.34 (0.92, 1.96)	1.15 (0.98, 1.34)	1.00 (0.69, 1.47)	1.10 (0.75, 1.60)
Q4 (0.03, 16.5)	1.72 (1.44, 2.06)	2.17 (1.42, 3.30)	1.66 (1.40, 1.98)	1.54 (1.01, 2.34)	1.62 (1.07, 2.45)

* Model 1: [binary indicators] sex, race/ethnicity (white vs. non-white), receipt of Medical Assistance (ever/never), smoking status (ever/never), body mass index (BMI, kg/m^2) (centered and centered-squared term), age category

** Model 2: Model 1 + year

§ Model 3: Model 1+ region. Regions were defined by county of residence as northeast (reference), southeast, central, southwest, and northwest [see **Figure 5**].

† Model 4: Model 1+ year and region

‡ Model 5: Model 1+ year, region, season (winter [reference], spring, summer, fall), distance to hospital/clinic, and contact time (date of case or control encounter minus the date of the first encounter in medical record, in days)

Table 3.7. Summary of model results from sensitivity analyses

	Model 1*	Model 2**	Model 3[§]
UNGD Metric	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**
Pad preparation, $1/m^2$			
Q1 (4.0×10^{-9}, 3.3×10^{-8})	(ref)	(ref)	(ref)
Q2 (3.3×10^{-8}, 6.2×10^{-8})	1.11 (0.97, 1.26)	1.15 (0.97, 1.36)	1.12 (0.97, 1.29)
Q3 (6.2×10^{-8}, 1.2×10^{-7})	1.40 (1.21, 1.63)	1.60 (1.32, 1.93)	1.49 (1.27, 1.75)
Q4 (1.2×10^{-7}, 2.5×10^{-5})	1.47 (1.23, 1.76)	1.63 (1.30, 2.05)	1.54 (1.27, 1.86)
Spud (drilling), $1/m^2$			
Q1 (1.9×10^{-10}, 2.6×10^{-9})	(ref)	(ref)	(ref)
Q2 (2.6×10^{-9}, 3.8×10^{-9})	0.97 (0.81, 1.15)	0.98 (0.79, 1.21)	0.97 (0.81, 1.17)
Q3 (3.8×10^{-9}, 4.7×10^{-8})	1.03 (0.86, 1.25)	1.04 (0.82, 1.32)	1.02 (0.84, 1.25)
Q4 (4.7×10^{-9}, 1.7×10^{-8})	0.95 (0.77, 1.18)	0.90 (0.69, 1.18)	0.91 (0.73, 1.14)
Stimulation, m/m^2			
Q1 (2.5×10^{-6}, 2.7×10^{-4})	(ref)	(ref)	(ref)
Q2 (2.7×10^{-4}, 5.6×10^{-4})	0.99 (0.82, 1.21)	1.04 (0.82, 1.33)	1.02 (0.83, 1.25)
Q3 (5.6×10^{-4}, 9.7×10^{-4})	1.32 (1.06, 1.64)	1.63 (1.25, 2.14)	1.49 (1.19, 1.87)
Q4 (9.7×10^{-4}, 0.2)	1.50 (1.19, 1.89)	1.79 (1.34, 2.40)	1.63 (1.28, 2.08)
Production, m^3/m^2			
Q1 (2.2×10^{-6}, 0.002)	(ref)	(ref)	(ref)
Q2 (0.002, 0.02)	0.88 (0.68, 1.13)	0.96 (0.70, 1.31)	0.95 (0.73, 1.23)
Q3 (0.02, 0.03)	1.00 (0.74, 1.36)	1.24 (0.85, 1.82)	1.17 (0.85, 1.61)
Q4 (0.03, 16.5)	1.34 (0.96, 1.86)	1.81 (1.19, 2.75)	1.67 (1.18, 2.35)

* Model 1: sex, race/ethnicity (white vs. non-white), receipt of Medical Assistance (ever/never), smoking status (ever/never), body mass index (BMI, kg/m^2), age category, year, region, season (winter [reference], spring, summer, fall), distance to hospital/clinic, and days since heart failure diagnosis (centered and centered-squared term)

**Model 2: same covariates as Model 1, but restricted by age categories (only ages 40-80), resulting in 8412 subjects with 10812 case events and control encounters. Model 2 also adjusts for contact time (date of case or control encounter minus the date of the first encounter in medical record, in days) instead of days since heart failure diagnosis, and includes two random intercepts for both individual subjects and for place type identifiers

[§] Model 3: same as Model 2, with adjustment for days since heart failure diagnosis instead of contact time

Table 3.8. Summary of model results from sensitivity analyses for 4:1 control encounter to case event matching

	Model 1*	Model 2**
UNGD Metric	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**
Pad preparation, $1/m^2$		
Q1 (3.9×10^{-9}, 3.2×10^{-8})	(ref)	(ref)
Q2 (3.2×10^{-8}, 6.1×10^{-8})	1.30 (1.14, 1.47)	1.49 (1.27, 1.74)
Q3 (6.1×10^{-8}, 1.2×10^{-7})	1.45 (1.26, 1.67)	1.77 (1.48, 1.96)
Q4 (1.2×10^{-7}, 4.8×10^{-4})	1.47 (1.23, 1.75)	1.82 (1.45, 2.28)
Spud (drilling), $1/m^2$		
Q1 (3.9×10^{-9}, 3.2×10^{-8})	(ref)	(ref)
Q2 (3.2×10^{-8}, 6.1×10^{-8})	1.29 (1.14, 1.47)	1.41 (1.20, 1.66)
Q3 (6.1×10^{-8}, 1.2×10^{-7})	1.45 (1.26, 1.67)	1.62 (1.25, 1.96)
Q4 (1.2×10^{-7}, 4.8×10^{-4})	1.47 (1.23, 1.75)	1.58 (1.26, 1.99)
Stimulation, m/m^2		
Q1 (2.4×10^{-6}, 2.6×10^{-4})	(ref)	(ref)
Q2 (2.6×10^{-4}, 5.3×10^{-4})	1.00 (0.83, 1.20)	1.08 (0.86, 1.37)
Q3 (5.3×10^{-4}, 9.2×10^{-4})	1.34 (1.09, 1.64)	1.54 (1.19, 2.00)
Q4 (9.2×10^{-4}, 1.5)	1.70 (1.36, 2.13)	2.05 (1.54, 2.72)
Production, m^3/m^2		
Q1 (1.9×10^{-6}, 0.002)	(ref)	(ref)
Q2 (0.002, 0.01)	0.84 (0.67, 1.05)	0.82 (0.61, 1.09)
Q3 (0.01, 0.03)	1.04 (0.78, 1.38)	1.06 (0.74, 1.52)
Q4 (0.03, 21.2)	1.47 (1.08, 2.02)	1.63 (1.08, 2.45)

*Model 1: 4:1 control to case sampling, 11774 subjects and 29051 case events or control encounters. Model 1 adjusts for: sex, race/ethnicity (white vs. non-white), receipt of Medical Assistance (ever/never), smoking status (ever/never), body mass index (BMI, kg/m^2) (centered and centered-squared term), age category, year, region, season (winter [reference], spring, summer, fall), and days since heart failure diagnosis (centered and centered-squared term)

**Model 2: same covariates as Model 1, but adjusting for contact time (centered and centered-squared term) instead of days since heart failure diagnosis

Table 3.9. Associations of UNGD activity metrics, by phase, with hospitalization for HF, from models with adjustment for time-dependent medication use, by class

	Model 1* (No medication adjustment)	Model 1+ Antihypertensive medications	Model 1 + Antihyperlipidemic medications	Model 1 + Anticoagulant medications
UNGD Metric	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**
Pad preparation, $1/m^2$				
Q1 (4.0×10^{-9}, 3.3×10^{-8})	(ref)	(ref)	(ref)	(ref)
Q2 (3.3×10^{-8}, 6.2×10^{-8})	1.13 (1.01, 1.41)	1.19 (1.01, 1.40)	1.19 (1.01, 1.41)	1.19 (1.01, 1.41)
Q3 (6.2×10^{-8}, 1.2×10^{-7})	1.64 (1.35, 1.98)	1.62 (1.34, 1.95)	1.63 (1.35, 1.97)	1.64 (1.35, 1.98)
Q4 (1.2×10^{-7}, 2.5×10^{-5})	1.70 (1.35, 2.13)	1.68 (1.34, 2.11)	1.69 (1.35, 2.13)	1.70 (1.35, 2.14)
Spud (drilling), $1/m^2$				
Q1 (1.9×10^{-10}, 2.6×10^{-9})	(ref)	(ref)	(ref)	(ref)
Q2 (2.6×10^{-9}, 3.8×10^{-9})	1.01 (0.82, 1.25)	1.02 (0.83, 1.25)	1.01 (0.82, 1.25)	1.01 (0.82, 1.25)
Q3 (3.8×10^{-9}, 4.7×10^{-8})	1.08 (0.85, 1.36)	1.08 (0.85, 1.35)	1.08 (0.85, 1.36)	1.08 (0.85, 1.36)
Q4 (4.7×10^{-9}, 1.7×10^{-8})	0.97 (0.74, 1.26)	0.97 (0.75, 1.26)	0.97 (0.75, 1.26)	0.97 (0.75, 1.26)
Stimulation, m/m^2				
Q1 (2.5×10^{-6}, 2.7×10^{-4})	(ref)	(ref)	(ref)	(ref)
Q2 (2.7×10^{-4}, 5.6×10^{-4})	1.03 (0.81, 1.32)	1.03 (0.81, 1.31)	1.03 (0.81, 1.31)	1.03 (0.81, 1.32)
Q3 (5.6×10^{-4}, 9.7×10^{-4})	1.57 (1.20, 2.05)	1.56 (1.19, 2.03)	1.56 (1.20, 2.05)	1.57 (1.20, 2.06)
Q4 (9.7×10^{-4}, 0.2)	1.80 (1.35, 2.41)	1.79 (1.34, 2.38)	1.80 (1.35, 2.41)	1.81 (1.35, 2.42)
Production, m^3/m^2				
Q1 (2.2×10^{-6}, 0.002)	(ref)	(ref)	(ref)	(ref)
Q2 (0.002, 0.02)	0.88 (0.65, 1.20)	0.88 (0.65, 1.20)	0.88 (0.65, 1.20)	0.88 (0.65, 1.20)
Q3 (0.02, 0.03)	1.11 (0.76, 1.62)	1.11 (0.76, 1.61)	1.11 (0.76, 1.62)	1.11 (0.76, 1.62)
Q4 (0.03, 16.5)	1.63 (1.08, 2.48)	1.62 (1.07, 2.45)	1.63 (1.08, 2.47)	1.63 (1.08, 2.48)

*Model 1: [binary indicators] sex, race/ethnicity (white vs. non-white), receipt of Medical Assistance (ever/never), smoking status (ever/never), body mass index (BMI, kg/m^2) as a centered and centered-squared term, age category, year, region, season (winter [reference], spring, summer, fall), distance to hospital/clinic, and observation time as a centered and centered-squared term (date of case or control encounter – date of first encounter in medical record, in days)

Table 3.10. Associations of UNGD activity metrics, by phase, with hospitalization for HF, from models with adjustment for related comorbidities

	Model 1* (No comorbidity adjustment)	Model 1+ chronic obstructive pulmonary disease	Model 1 + Type II diabetes	Model 1 + Coronary artery disease	Model 1 + Myocardial infarction	Model 1 + Hypertension	Model 1 + chronic kidney disease	Model 1 + Valve disease
UNGD Metric	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**	**Odds ratio (95% CI)**
Pad preparation								
Q1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
Q2	1.19 (1.01, 1.41)	1.20 (1.01, 1.41)	1.19 (1.01, 1.41)	1.21 (1.02, 1.42)	1.20 (1.01, 1.42)	1.19 (1.01, 1.41)	1.19 (1.00, 1.40)	1.19 (1.01, 1.41)
Q3	1.64 (1.35, 1.98)	1.64 (1.36, 1.98)	1.64 (1.35, 1.98)	1.66 (1.37, 2.01)	1.66 (1.37, 2.01)	1.63 (1.35, 1.97)	1.64 (1.35, 1.98)	1.64 (1.36, 1.98)
Q4	1.70 (1.35, 2.13)	1.70 (1.35, 2.14)	1.71 (1.36, 2.15)	1.74 (1.38, 2.19)	1.72 (1.36, 2.16)	1.68 (1.34, 2.12)	1.70 (1.35, 2.14)	1.70 (1.35, 2.14)
Spud (drilling)								
Q1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
Q2	1.01 (0.82, 1.25)	1.01 (0.82, 1.24)	1.01 (0.82, 1.25)	1.02 (0.82, 1.26)	1.01 (0.82, 1.25)	1.01 (0.82, 1.25)	1.01 (0.81, 1.24)	1.01 (0.82, 1.25)
Q3	1.08 (0.85, 1.36)	1.07 (0.85, 1.35)	1.07 (0.85, 1.36)	1.08 (0.85, 1.37)	1.07 (0.85, 1.35)	1.07 (0.85, 1.35)	1.06 (0.84, 1.35)	1.08 (0.85, 1.36)
Q4	0.97 (0.74, 1.26)	0.97 (0.74, 1.26)	0.97 (0.74, 1.26)	0.98 (0.75, 1.28)	0.97 (0.74, 1.26)	0.96 (0.74, 1.26)	0.96 (0.74, 1.26)	0.97 (0.75, 1.26)
Stimulation								
Q1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
Q2	1.03 (0.81, 1.32)	1.03 (0.81, 1.32)	1.03 (0.81, 1.32)	1.04 (0.82, 1.33)	1.02 (0.80, 1.30)	1.03 (0.81, 1.31)	1.06 (0.83, 1.35)	1.03 (0.81, 1.32)
Q3	1.57 (1.20, 2.05)	1.57 (1.20, 2.05)	1.57 (1.20, 2.06)	1.60 (1.22, 2.10)	1.54 (1.18, 2.03)	1.56 (1.19, 2.04)	1.61 (1.23, 2.12)	1.57 (1.20, 2.05)
Q4	1.80 (1.35, 2.41)	1.80 (1.35, 2.41)	1.81 (1.36, 2.42)	1.84 (1.37, 2.47)	1.79 (1.34, 2.40)	1.79 (1.34, 2.40)	1.87 (1.39, 2.49)	1.81 (1.35, 2.42)
Production								
Q1	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)	(ref)
Q2	0.88 (0.65, 1.20)	0.88 (0.65, 1.20)	0.88 (0.65, 1.20)	0.89 (0.65, 1.22)	0.87 (0.63, 1.18)	0.87 (0.64, 1.18)	0.90 (0.66, 1.23)	0.87 (0.64, 1.20)
Q3	1.11 (0.76, 1.62)	1.11 (0.76, 1.62)	1.11 (0.76, 1.63)	1.13 (0.77, 1.66)	1.10 (0.75, 1.61)	1.10 (0.75, 1.60)	1.14 (0.78, 1.68)	1.11 (0.76, 1.62)
Q4	1.63 (1.08, 2.48)	1.64 (1.08, 2.48)	1.64 (1.08, 2.48)	1.69 (1.11, 2.58)	1.63 (1.07, 2.48)	1.61 (1.06, 2.43)	1.70 (1.12, 2.59)	1.63 (1.08, 2.48)

*Model 1: [binary indicators] sex, race/ethnicity (white vs. non-white), receipt of Medical Assistance (ever/never), smoking status (ever/never), body mass index (BMI, kg/m^2) as a centered and centered-squared term, age category, year, region, season (winter [reference], spring, summer, fall), distance to hospital/clinic, and observation time as a centered and centered-squared term (date of case or control encounter – date of first encounter in medical record, in days)

Figure 3.3. Map of subjects included in analysis from 2008 to 2015, by quartile of the assigned UNGD spud activity metric.

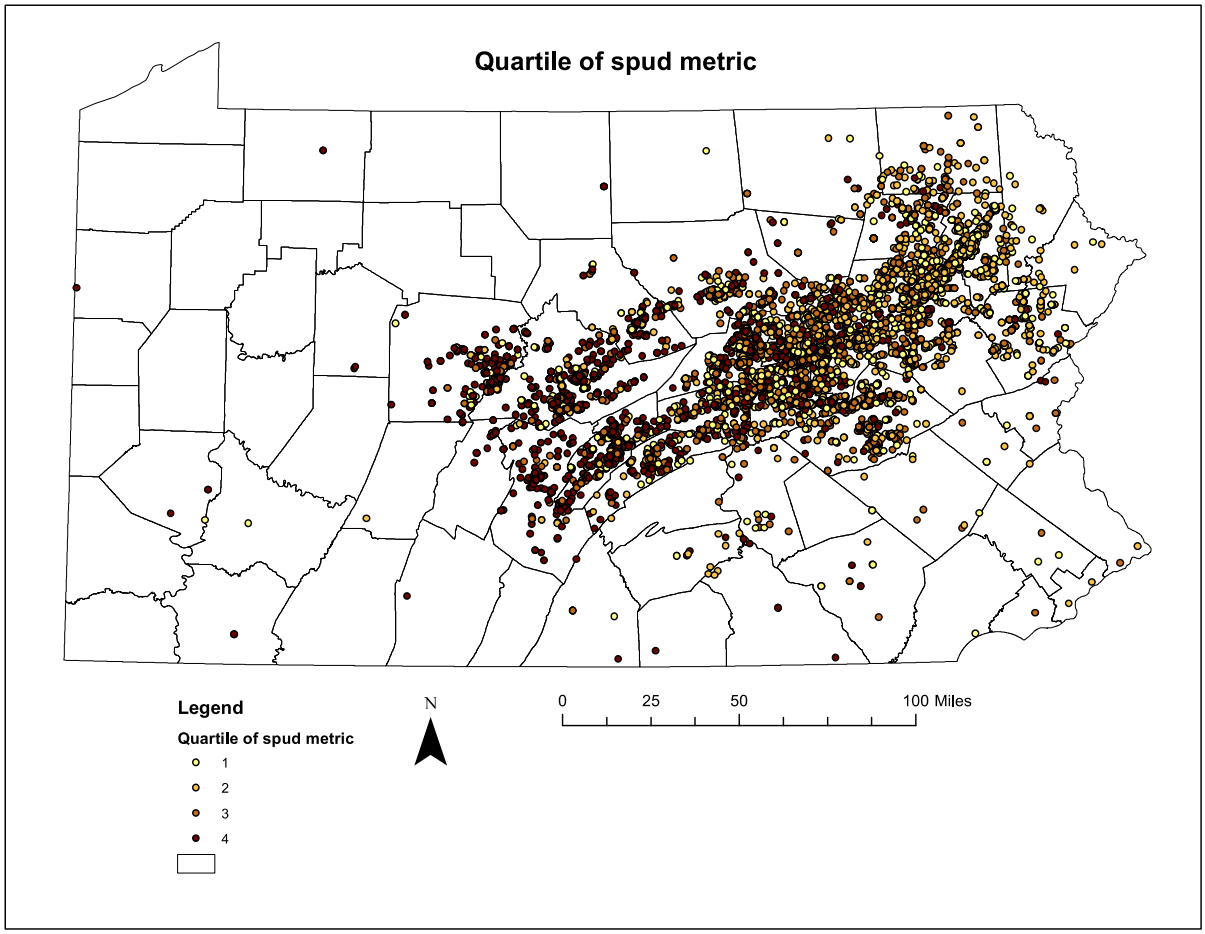


Figure 3.4. Map of subjects 2008-2015, by quartile of the UNGD production activity metric.

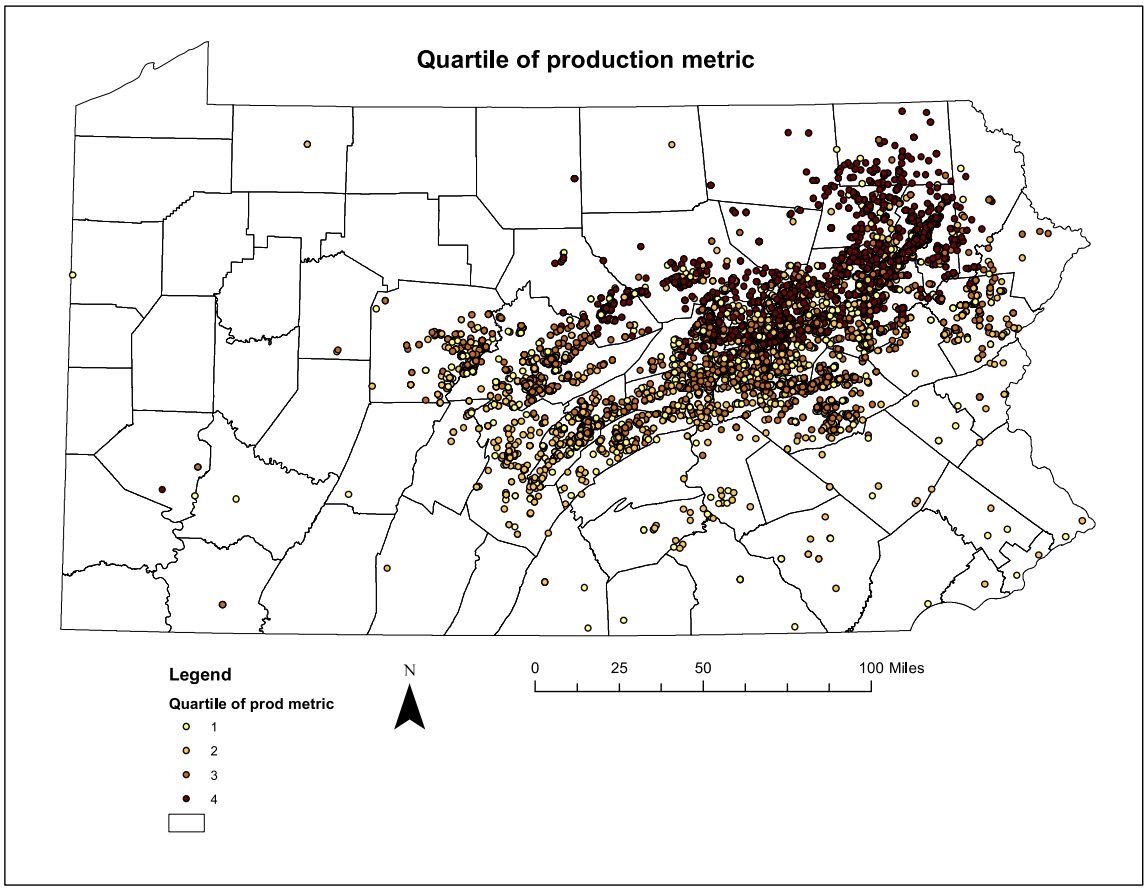


Figure 3.5. Map of the distribution of the regional indicator variable, by county of residence. For the analysis, northeast was the reference region.

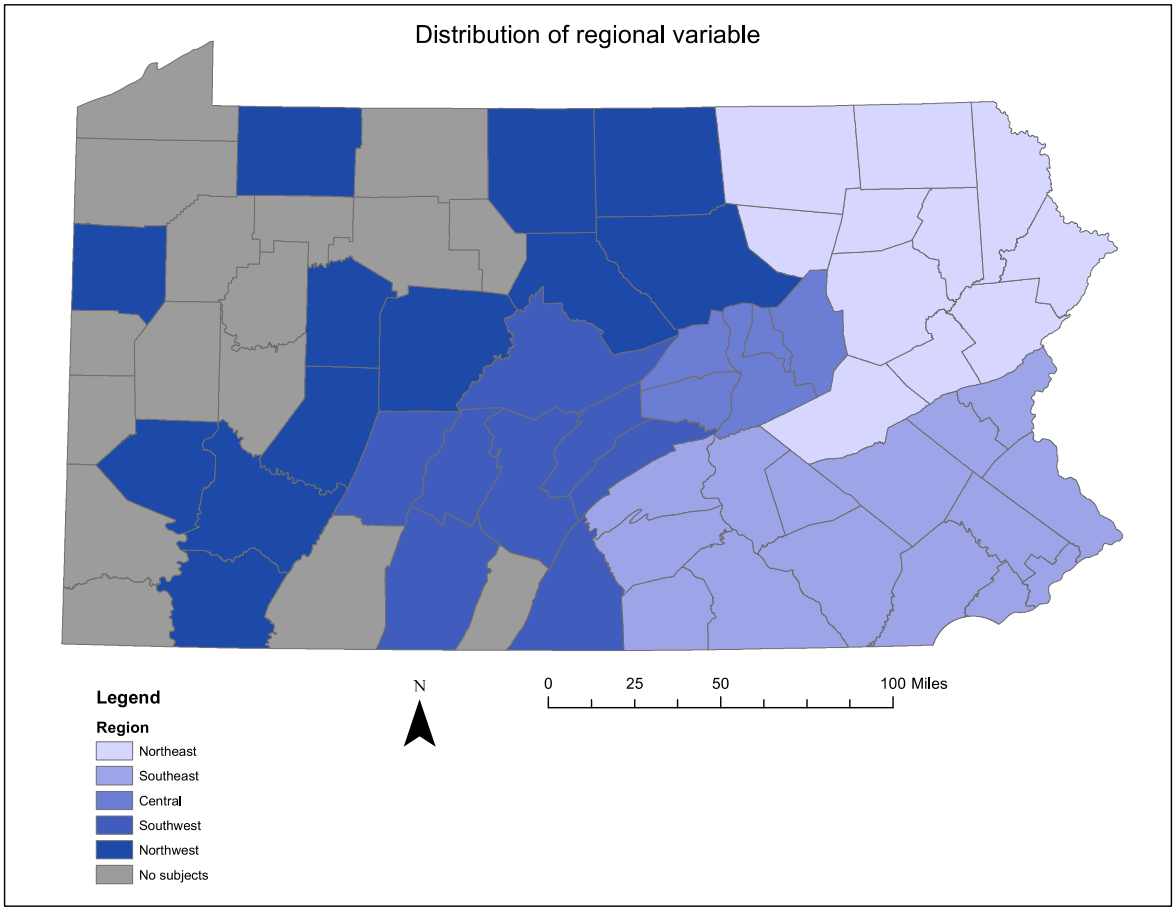
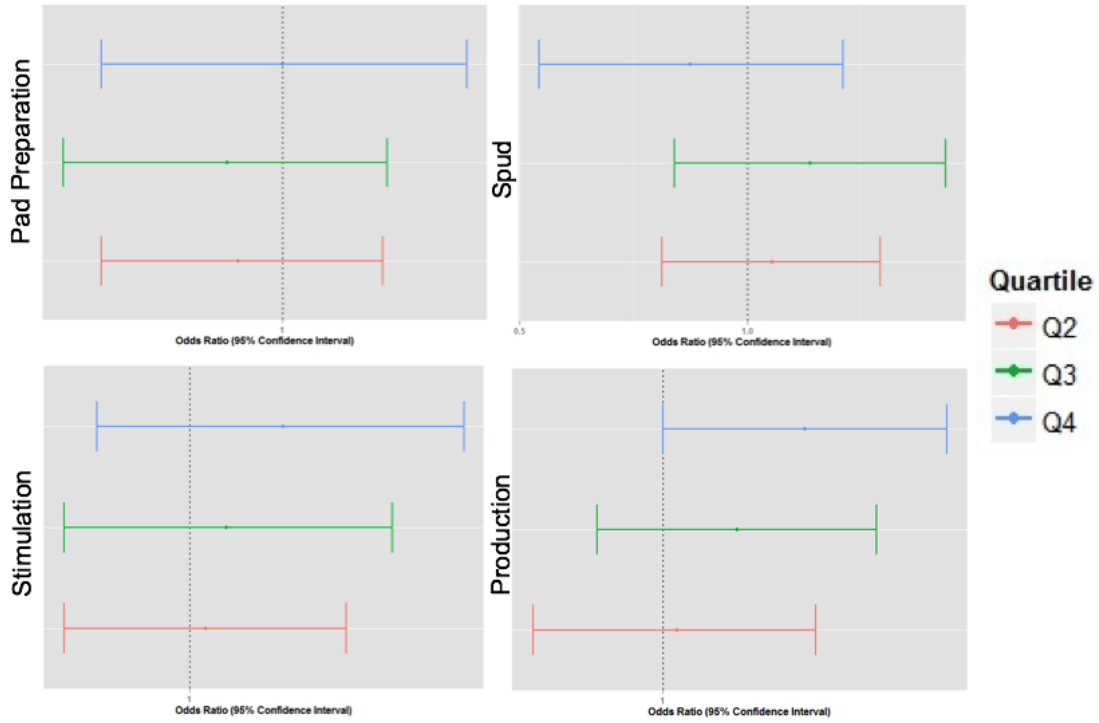


Figure 3.6. Forest plots of adjusted odds ratio for heart failure hospitalization for each metric from negative exposure control* models.



*Negative exposure control models limited our analyses to years 2008 and 2009, however UNGD activity metrics were assigned using data from 2014 and 2015

Chapter 4: Environmental factors associated with B-type Natriuretic Peptide levels among heart failure patients in Pennsylvania

Authors: Tara P. McAlexander, MPH[1], Karen Bandeen-Roche, PhD, MS[2,3], Jessie P. Buckley, PhD, MPH[1,4], Jonathan Pollak, MS[1], Erin D. Michos, MD, MHS[3], John William McEvoy, MBBCh, MHS[3,5], Brian S. Schwartz, MD, MS[1,3,4]

Affiliations
[1] Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States
[2] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States
[3] Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States
[4] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States
[5] National Institute for Preventive Cardiology, National University of Ireland, Galway, Ireland

Corresponding Author: Brian S. Schwartz, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Room W7041, Baltimore, Maryland 21205, email bschwar1@jhu.edu
Running title: Environmental factors and B-type natriuretic peptide

Acknowledgements

This research was supported by funding from the National Institute of Environmental Health Sciences (NIEHS) training grant # 5T32ES007141-32 (T McAlexander) and NIEHS grant ES023675-01 (PI: B Schwartz). Additional support was provided by the Degenstein foundation for the compiling of well data. The authors would like to thank Joseph DeWalle, Dione Mercer, and staff of the Geisinger Environmental Health Institute for their assistance in obtaining data essential for this study.

Competing financial interest disclosure: The authors declare that they have no competing financial interests.

4.1 ABSTRACT

Background: B-type natriuretic peptide (BNP) has been increasingly used in clinical settings to aid in the diagnosis and management of heart failure. Given associations between unconventional natural gas development (UNGD) activity and heart failure hospitalization observed in **Chapter 3**, we sought to understand whether UNGD activity was similarly associated with BNP. To date, epidemiologic studies have not yet evaluated UNGD activity and other environmental factors in relation to biomarkers of heart failure in a large patient population. BNP levels may offer advantages over heart failure hospitalization in several ways: 1) it may be an earlier marker of exacerbating heart failure; 2) it is more routinely measured and would not be biased, for example, by health care system programs to increase outpatient management of heart failure and reduce hospitalization; and 3) it is mechanistically and pathophysiologically important to heart failure.

Objectives: To estimate the association between metrics of UNGD activity, related environmental factors, and BNP concentrations in a large patient population in Pennsylvania.

Methods: We obtained Geisinger electronic health record (EHR) data for patients with heart failure diagnoses and laboratory orders for BNP. We generated quantitative estimates of UNGD activity in the 30 days prior to the BNP laboratory order date, by phase (i.e., pad preparation, spud drilling, stimulation, and production), with a one-day lag. We also generated metrics of community socioeconomic deprivation (CSD), residential greenness, and distances to major and minor roads as well as comorbidities and medication use from the EHR. We evaluated associations of these variables with BNP levels dichotomized at greater than or equal to 400 pg/mL (vs. lower), a common

clinically-used cutoff, using logistic regression and generalized estimating equations (GEE).

Results: We identified 6501 laboratory values for BNP among 3938 subjects with heart failure in the years 2008-2015. The mean (SD) BNP concentration for the 6501 laboratory measures was 478.8 pg/mL (637.5). Adjusted associations (odds ratio [OR] and 95% confidence interval [CI]) of UNGD activity with BNP levels ≥ 400 pg/mL were 1.36 (1.08, 1.71), 1.42 (1.05, 1.93), and 1.52 (1.07, 2.17), for the 2nd, 3rd, and 4th quartiles of UNGD production compared to the first, respectively. We did not observe any consistent associations for other UNGD metrics with BNP ≥ 400 pg/mL. In sensitivity analyses using linear regression and *ln*-transformed BNP, the geometric mean (95% CI) BNP was 1.28 (1.15, 1.43), 1.30 (1.12, 1.51), and 1.39 (1.17, 1.65) times greater for the 2nd, 3rd, and 4th quartiles of UNGD production compared to the 1st quartile, after adjusting for covariates. We did not see consistent associations with other UNGD metrics (pad preparation, stimulation) or with CSD, NDVI, and distance to major roads. The second quartile of distance to minor roads (vs. the first), however, showed a geometric mean ratio of 0.90 (95 % CI: 0.81, 0.99).

Conclusions: Higher levels of UNGD production activity were associated with higher BNP in both continuous and dichotomous outcome models. The exposure-effect relation became apparent with increasing covariate control. These findings are consistent with the results of **Chapter 3**, which showed exposure-effect relations between UNGD activity metrics and heart failure hospitalization; however, unlike **Chapter 3**, exposure-effect relations in this chapter were only present for the UNGD production metric. There are several possible explanations for the discordant results between this chapter and those in **Chapter 3**, having to do with the timing of BNP measurements, the fact that the production metric is the only one that is cumulative (the others have defined beginnings

and endings), and possibly unique aspects of production activity exposures. These are discussed herein.

4.2 Introduction

Clinical measures from EHRs provide a valuable data resource for understanding the epidemiology of prevalent and burdensome diseases for which persons must seek and receive health care. Heart failure is a disease with a high public health burden, and heart failure patients provide many clinical measures when in contact with the health care system. In particular, blood levels of B-type natriuretic peptide (BNP) are frequently used in a clinical setting as diagnostic and prognostic markers among heart failure patients [28, 235, 236], even those who are asymptomatic [237]. At the population-level, recent epidemiologic studies underscore the importance and utility of blood BNP measures obtained from EHRs in defining clinical management and understanding patient survival and risk prediction [31, 238].

In vivo, BNP is a vasodilator that is released by cardiac myocytes in response to increases in volume and pressure in the left ventricle [28]. The release of BNP into the bloodstream leads to diuresis, natriuresis, vasodilation, and the inhibition of renin, aldosterone, and fibrosis [235]. It is cleared from the blood via both receptor-mediated mechanisms and by endopeptidase enzymes [107, 235]. The reported half-life of BNP in the blood varies from 3.9 to 22.6 minutes [107], which allows clinicians to utilize blood BNP levels for diagnosis of acute heart failure [235]. Specifically, blood BNP levels can be used to *exclude* acute heart failure (< 30 pg/mL), to *identify* acute heart failure (> 100 pg/mL), to screen asymptomatic patients for heart failure in an outpatient setting (> 20 pg/mL), and to *rule in* acute heart failure for persons experiencing acute dyspnea (> 400 pg/mL) [235]. For these reasons, many heart failure patients, and patients suspected to

have heart failure, will have multiple BNP laboratory measurements over the duration of their care. A patient's blood BNP concentration can vary over time and can be impacted by certain medications (e.g., diuretics, angiotensin-II inhibitors) and various comorbid conditions (e.g., chronic kidney disease, obesity) [235]. Growing evidence links the circulation of BNP and related cardiac biomarkers to metabolic dysfunction and dysregulation of glucose and blood lipids, providing a potential mechanism for the interplay between heart failure, hypertension, and metabolic disorders [239, 240]. Because of these complexities, epidemiologic studies of BNP need to consider a suite of medications and comorbid conditions, many of which are documented in EHRs.

4.2.1 Epidemiology and biologic rationale for associations between environmental factors and BNP

There is substantial epidemiologic evidence that environmental exposures adversely affect cardiac health. Air pollution, and particulate matter less than 2.5 microns in diameter ($PM_{2.5}$) in particular, is consistently associated with increased cardiovascular mortality and morbidity in epidemiologic studies [128]. These associations also apply to heart failure patients, where multiple epidemiologic studies have found short-term associations between criteria air pollutants, such as $PM_{2.5}$, NO_x, and ozone, and hospitalization among heart failure patients [26]. The biologic mechanisms underlying these associations are complex, but candidate pathways responsible for these associations include systemic inflammation [134] and endothelial injury [241, 242], leading to vasoconstriction [243]. Associations between air pollution and elevations in blood pressure are also well-documented in epidemiologic studies [244]. In addition to air pollution, there is evidence that both psychosocial stress and noise exposure are directly associated with endothelial dysfunction [245, 246]. Increasing evidence implicates environmental noise exposure as an independent risk factor for

cardiovascular disease [132, 247]. To the best of our knowledge, however, there are only a few epidemiologic studies of exposure to environmental factors (e.g., air pollution, noise) and psychosocial stress in relation to BNP [124-126, 248, 249]. Notably, as discussed throughout this dissertation, each of these exposures can be caused by UNGD. Lastly, although there is evidence that community socioeconomic deprivation is independently associated with cardiovascular disease [158, 250], and that residential greenness is associated with improved cardiovascular outcomes [52, 151, 251], reduced odds of heart failure [52], we have not been able to identify any epidemiologic studies that evaluate associations of residential greenness, proximity to major and minor roads, or community socioeconomic deprivation (CSD) in relation to BNP.

4.3 OBJECTIVE

Heart failure is a heterogeneous disease and associations between environmental factors and hospitalization for heart failure likely reflect multiple biologic pathways that lead a subject to seek care and become hospitalized. Because BNP levels in heart failure patients are associated with prognosis and severity [30], and because of our observed associations between UNGD activity and hospitalization among heart failure subjects (**Chapter 3**), we hypothesized that UNGD activity might also be associated with BNP measurements as a biomarker of early biologic effect in the association between UNGD activity and exacerbation of heart failure. Although we do not believe that BNP is a unique marker of exposure to UNGD activity, examining associations between UNGD activity and BNP allowed us to see if this biologically plausible association was present in a large population of heart failure subjects, which can inform understanding of the biologic pathways through which environmental factors contribute to heart failure outcomes. By contrasting associations of UNGD activity separately with heart failure hospitalization and BNP levels, we hypothesized that we

could learn about key aspects of how UNGD may affect heart failure, the utility of EHR data for environmental epidemiology, and possibly gain mechanistic insight. BNP may be an earlier marker of exacerbating heart failure, it is more routinely measured, and it would not be biased, for example, by health care system programs to increase outpatient management of heart failure and reduce hospitalization. The objective of this study was to estimate the association between metrics of UNGD activity, related environmental factors, and laboratory measures of BNP in a large patient population in Pennsylvania.

4.4 METHODS

4.4.1 STUDY POPULATION AND LABORATORY MEASURES

We identified 16,098 individuals who were evaluated at Geisinger and had at least one ICD-9 code for heart failure (428.x) between 2003 - 2015 included in their medical record. We applied the subject eligibility criteria that is described in **Chapter 2, Figure 2.1** (i.e., requiring a residential address in Pennsylvania, excluding subjects with missing demographic information or with diagnoses of endocardial fibroelastosis or congenital heart anomalies) to the laboratory orders we identified, so that 13,183 subjects were eligible for potential selection. We identified 7224 subjects who had at least one laboratory procedure code order for “83880.01” or “83880.03” for B-type natriuretic peptide (see **Chapter 2, Table 2.1**). Because we wanted the laboratory measurements from this study population to coincide with the timing of UNGD activity in Pennsylvania, we restricted our study to the years 2008 - 2015, leaving us with 4893 potentially eligible subjects with BNP measures for this analysis. We then excluded laboratory measures from 922 subjects whose BNP laboratory measures were recorded before their first diagnosis for heart failure and 33 subjects who did not have a BNP value despite the laboratory order; resulting in 3938 subjects in this analysis, all of whom

were over the age of 18 years at the time of each laboratory measurement. We further limited the number of laboratory measurements to be included in the model to be no more than five (**Table 4.1**), so that individuals with many observations, many of which were clustered in time and possibly indicating acute exacerbation, would not have undue influence in our model. We also restricted these BNP orders to be at least 90 days before or after the adjacent BNP order. We created variables to indicate the number of days before or after a subject's BNP order and the date of hospitalization for heart failure, if present.

4.4.2 Covariate assignment for EHR variables

The Geisinger EHR allowed for identification of demographic characteristics such as race/ethnicity and sex. For each subject in this analysis, we also created time-varying variables based on each laboratory order date, including age at time of event, receipt of Medical Assistance (ever vs. never), smoking status (ever vs. never), and body mass index (BMI, mg/kg^2). We used each subject's most recently available height and weight data prior to the date of the laboratory order to calculate BMI. For a small (i.e., < 5%) subset of subjects, height and weight data were missing from the medical record. We imputed BMI for these subjects using multiple imputation based on: age at laboratory order date (centered), sex, race/ethnicity, smoking status, and receipt of Medical Assistance. We assigned smoking status in the same manner as it was assigned in **Chapter 3**.

We created variables for a subject having comorbid conditions associated with BNP levels or clearance of BNP from the blood by identifying at least two ICD-9 diagnosis codes for each condition from inpatient, outpatient, ER visits and medication orders. These conditions were dichotomized as ever vs. never, considering all available EHR data for the subject prior to the date of the laboratory order. The comorbid

conditions that we deemed important for this analysis included: hypertension (401.9), type 2 diabetes (250.x), myocardial infarction (410.x), chronic obstructive pulmonary disease (496.x), heart valve disorders and disease (V43.3, 424.x, 396.9, 391.1, 392.0, 390, 395.9, or 421.9), coronary artery disease (414.01), and chronic kidney disease (585.x).

We used stringent criteria to define our medication variables. We used the Medi-Span Generic Product Identifier Therapeutic Classification System [191] to identify medication orders by the following medication classes: antihypertensive, antihyperlipidemic, and anticoagulant medications. We created a binary variable for whether or not a subject had a valid medication order that encompassed the date of the laboratory order. Lastly, we calculated the Charlson index, a 17-item morbidity scale, at the date of the laboratory order. More than a quarter of subjects (n = 1082, 27.5 %) did not have all of the necessary information in their medical record to generate this 17-item scale, so we imputed Charlson index for these subjects using multiple imputation based on: age at laboratory order (centered), sex, race/ethnicity, smoking status, and receipt of Medical Assistance.

4.4.3 Environmental and community variables

We assigned environmental and community variables in the same manner as described in **Chapter 3**. For each subject, we generated several place-based variables that we hypothesized might be relevant to heart failure and BNP. First, we assigned each subject a community type based on latitude and longitude coordinates geocoded from the residential address listed in the EHR (census tracts [cities], boroughs, and townships), as previously reported [13, 190, 192]. We obtained information from the 2010-2014 US Census American Community Survey [198] to generate a standardized index of community socioeconomic deprivation (CSD) at each community type [252].

Similar to our methods in **Chapter 3**, CSD was categorized into quartiles for the analysis.

4.4.3.1 UNGD ACTIVITY

Similar to the methods used in **Chapter 3**, we obtained the dates and locations of unconventional natural gas wells drilled and assigned UNGD activity for a 30-day duration and one-day lag prior to date of laboratory order (see **Chapter 2, Section 5.1** for detail on determining UNGD activity durations and lags). We did this for four phases of UNGD activity: pad preparation, spud (drilling), stimulation, and production of the well. These methods are consistent with previous epidemiological work on UNGD in this region and patient population [15, 190]. Because we wanted to identify phase-specific associations between 30-day UNGD activity and BNP measures, we created separate models for each of these four UNGD metrics.

4.4.3.2 NORMALIZED DIFFERENCE VEGETATION INDEX (NDVI) ASSIGNMENT

To evaluate the association between residential greenness and BNP, we first obtained data from the NASA MODIS satellite for the 16-day periods of annual maximum greenness for each calendar year from 2008-2015. We assigned the normalized difference vegetation index (NDVI), a measure of greenness, to subjects based on the NDVI values in the 1250 m x 1250 m grid surrounding their residential address, as has been done previously [57] (see **Chapter 2, Section 2.5** for further detail on this data source). Maximum greenness values were assigned by the calendar year of each BNP laboratory order.

4.4.3.3 DISTANCE TO MAJOR & MINOR ROAD AND NEAREST HOSPITAL OR CLINIC

We obtained road network data from the Federal Highway Administration for both major and minor road classifications. Using ArcGIS 10.4, we calculated the Euclidian distance (in meters) from each subject's residential address to both the nearest major

and minor roads using the Generate Near Table function. Similarly, we obtained the geographic locations of Geisinger's clinics and hospitals in Pennsylvania, and we calculated the Euclidian distance (in meters) of each subject's residential address to the nearest Geisinger hospital or clinic.

4.4.4 STATISTICAL ANALYSIS

We first examined the distribution of laboratory orders by person (see **Chapter 6, Figures 6.3-6.12**) to understand how subjects' BNP values changed over time. We suspected that laboratory order setting (i.e., inpatient vs. outpatient) could be a confounder in our main analysis, so we examined subject characteristics by laboratory order setting (**Table 4.2**). To do so, we used only one measure per person (i.e., each subject's first BNP laboratory measure) and conducted analysis of variance (ANOVA) F-tests for continuous variables, and chi^2 tests for categorical variables, to assess if any of these variables differed by subjects whose first BNP values were obtained in either an inpatient or outpatient setting. We then examined differences in mean BNP values across all selected BNP measures per subject by categorical factors using GEE models (**Table 4.3**).

We dichotomized BNP measurements by clinical guidelines (i.e., whether a BNP value was < 400 pg/mL or ≥ 400 pg/mL, a common definition of acute heart failure [235]), and we examined descriptive statistics of the study population divided in this way (**Table 4.4**). We decided to dichotomize BNP for this analysis because this is a standard clinical approach to evaluate if acute dyspnea in individuals with heart failure is due to heart failure exacerbation [235], and because BNP evidences considerable intra-individual and inter-individual variation that may not be biologically relevant to this

analysis [253], suggesting that modeling BNP on a continuous scale may be less meaningful.

Since we standardized the minimum time period between adjacent laboratory measurements from each subject and limited the number of values per subject to no more than five, we were able to utilize logistic regression with generalized estimating equations to estimate the adjusted associations of UNGD metrics, CSD, NDVI, and distance to major/minor roads with the odds of high (≥ 400 pg/mL) vs. low (< 400 pg/mL) BNP values. We examined the associations between environmental factors in separate models because the metrics for UNGD, CSD, NDVI, and distance to major/minor roads evidenced non-positivity [254, 255] when included together, and our conceptual framework did not suggest that these were potential confounding variables (we were mainly interested in their main effect associations). In each of these models, we implemented an exchangeable correlation matrix [256].

We first built our models for BNP based on our *a priori* knowledge of important factors such as: age, sex, race/ethnicity, smoking status, and receipt of Medical Assistance. We then evaluated how these associations changed after adjusting for medication use (antihyperlipidemic and anticoagulant medications), comorbidities as measured by the Charlson index, BMI, and inpatient vs. outpatient setting. For continuous variables such as the Charlson index and BMI, we first centered each of these variables and evaluated them as both centered and centered-squared terms to account for nonlinearity. Additionally, we evaluated several variables related to the timing of subjects' BNP measurements: we suspected that the duration of the subjects' heart failure might be important for their BNP measures, so we evaluated duration, in days, from each subjects' first heart failure ICD-9 code to the date of the BNP laboratory value. Since these variables concerning time were measured as continuous variables (i.e. days), we centered these variables and evaluated them separately in models as

both centered and centered-squared terms. In model building, we retained variables if they resulted in more than a 5% change in the main effect of each environmental variable on BNP levels. We display this process of model building in **Table 4.5**. We evaluated model fit by examining the plots of model residuals and by goodness of fit tests [257].

We first adjusted for sex (female vs. male), race/ethnicity (nonwhite [including some Hispanic subjects] vs. white), smoking status (ever vs. ever), age at time of BNP value (centered), laboratory setting (inpatient vs. outpatient), receipt of Medical Assistance (ever vs. never up to time of BNP value), Charlson index (centered and centered squared), BMI (centered and centered squared), and chronic kidney disease (**Model 1**). We deemed it important to include chronic kidney disease in our base model because the condition directly impacts BNP levels, however we did not observe any effect modification of the associations between UNGD and BNP by chronic kidney disease [258]. **Model 2** additionally adjusted for anticoagulant and antihyperlipidemic medications, and **Model 3** included adjustment for duration of heart failure. Finally, we evaluated additional adjustment for year (**Model 4**), region (**Model 5**), and both region and year together (**Model 6**). These adjustments were important because the UNGD production metric considers total volume of natural gas produced, and so it only increases in value over time (i.e., it is the only UNGD activity metric that is cumulative over time – during our study period, once production started in a given well it did not stop, in contrast to the other three activity metrics, which each have defined starts and ends). We deemed adjustment for region important as well, as there were spatial patterns in the environmental metrics considered, and we wanted to adjust for to estimate more precisely the association between UNGD activity and odds of a BNP value $\geq$ 400 pg/mL.

4.4.4.1 SENSITIVITY ANALYSES

We conducted several sensitivity analyses. First, to evaluate whether our initial models could be biased from individuals who had more than one BNP laboratory value in the analysis, we randomly selected one BNP value per person and repeated models (**Table 4.6**). Second, we repeated models with inverse probability weights generated from the odds of being selected from the 13,183 eligible subjects (details on inverse probability weighted analysis in **Chapter 2, Section 8.2**). We evaluated Inverse probability weighted models in two ways: first, we utilized all generated weights (**Table 4.7, Model 1**), and second, we truncated weights at the 99th percentile and evaluated a model that accounted for truncated inverse probability weights (**Table 4.7, Model 2**). A final sensitivity analysis used linear regression to model *ln*-BNP as a continuous outcome (**Table 4.8**). Because the distribution of raw BNP values was right-skewed (**Figure 4.1a**), we natural log transformed BNP values (**Figure 4.1b**) so that we did not violate normality assumptions in this model.

4.5 RESULTS

4.5.1 DESCRIPTION OF STUDY SUBJECTS, BNP VALUES, AND UNGD METRICS

We identified 3938 subjects with 6501 laboratory measurements for BNP. These subjects had a mean (SD) age of 71.7 (12.0) years and were comprised of 2097 (53.2%) males and 1841 (46.7%) females. Study subjects were predominantly white (n = 3834, 97.4%) and 408 (10.4%) had ever received Medical Assistance. The mean (SD) BNP value among these 6501 laboratory measurements was 478.8 (637.5) pg/mL with a median value of 258 pg/mL. The majority (62.0%) of the 3938 subjects had only one BNP value (**Table 4.1**), however we included up to a maximum of five BNP measures for the individuals who had this information (n = 128). We also observed spatial patterns in

the environmental metrics we evaluated (**Figures 4.2a-e**), with clearer spatial patterns in the UNGD metrics (**Figures 4.2b-e**) than in the distribution of NDVI quartiles across study subjects (**Figure 4.2a**).

4.5.2 UNADJUSTED ASSOCIATIONS

We examined the setting of each subjects' first BNP laboratory order and determined that 2235 (56.8%) were obtained in an inpatient setting and 1703 (43.3%) in an outpatient setting. We used these first laboratory measures to present descriptive statistics of the study population (**Table 4.2**) by inpatient vs. outpatient setting. Subjects whose laboratory orders were obtained from an outpatient (vs. inpatient) setting had a higher usage of antihypertensive ($p < 0.001$), antihyperlipidemic ($p < 0.001$), and anticoagulant ($p = 0.003$) medications. Subjects whose laboratory orders were obtained from an inpatient (vs. outpatient) setting were more likely to have a diagnosis of chronic kidney disease ($p < 0.001$), hypertension ($p < 0.001$), myocardial infarction ($p = 0.02$), chronic obstructive pulmonary disease ($p < 0.001$), and type 2 diabetes ($p < 0.001$). Those who had their laboratory measures obtained from an inpatient (vs. outpatient) setting were slightly younger (mean age = 70.7 years vs. 72.2 years, $p < 0.001$) and lived slightly closer to the nearest hospital or clinic ($p = 0.009$) (**Table 4.2**).

We next evaluated associations of several comorbidities, medications, and demographic factors with continuous BNP levels using GEE models of *ln*-BNP (**Table 4.3**) across all 6501 laboratory measures from the 3938 subjects. We observed higher BNP levels in inpatient (vs. outpatient, $p < 0.001$) settings. Individuals receiving Medical Assistance had, on average, lower BNP levels than those who never received Medical Assistance ($p < 0.001$). Laboratory orders with the presence of comorbidities such as hypertension, myocardial infarction, valve disease, chronic obstructive pulmonary

disease, type 2 diabetes, and coronary artery disease had higher mean BNP values than those who did not ($p < 0.001$, $p < 0.001$, $p < 0.001$, $p = 0.04$, $p = 0.01$ and $p = 0.002$, respectively).

There were no differences in mean BNP values by quartile of CSD or the pad preparation, stimulation, or production UNGD metrics. In contrast, mean BNP did differ by quartiles of the spud metric ($p = 0.01$), but not in a clear exposure-effect pattern. Mean BNP values declined across quartiles of NDVI (p for trend = 0.06).

We evaluated associations of selected variables with BNP values divided at 400 pg/mL as previously described. There was a lower proportion of subjects in the 4th quartile of NDVI with high (vs. low) BNP values ($p = 0.08$, **Table 4.4**). Subjects with BNP values $\geq$ 400 pg/mL (vs. lower) were slightly older (mean age of 73.7 years vs. 71.0 years, $p < 0.001$), had higher Charlson index scores (9.7 vs. 8.6, $p < 0.001$), and a higher prevalence of such comorbidities as chronic kidney disease ($p < 0.001$), previous myocardial infarction ($p < 0.001$), valve disease ($p < 0.001$), and hypertension ($p = 0.007$) (**Table 4.4**). Subjects with high (vs. low) BNP values lived closer to minor roads (1384 meters vs. 1615 meters, $p = 0.003$).

4.5.3 Adjusted associations of UNGD and other environmental factors with BNP

We next used logistic regression with generalized estimating equations (GEE) with an exchangeable correlation matrix to evaluate associations between UNGD activity, NDVI, CSD, and distance to major/minor roads with high (vs. low) BNP values (**Table 4.5**). There were no associations (OR [95% CI]) of UNGD metrics with high (vs. low) BNP values in **Model 1,** with the exception of the 2nd quartile of the production metric (1.20 [1.04, 1.37]). After adjusting for anticoagulant and antihyperlipidemic medications, the 2nd (vs. 1st) quartiles of both the spud metric and distance to minor roads were associated with BNP (**Table 4.5, Model 2**). Additional adjustment for

duration of heart failure did not substantially change any associations (**Table 4.5, Model 3**). Finally, we evaluated additional adjustment for year (**Table 4.5, Model 4**), region (**Table 4.5, Model 5**), and both region and year together (**Table 4.5, Model 6**). In the final, fully-adjusted model (**Table 4.5, Model 6**), there was an exposure-effect relation (OR [95% CI]) of UNGD production activity with BNP across UNGD quartiles of 1.36 (1.08, 1.71), 1.42 (1.05, 1.93), and 1.52 (1.07, 2.17) for the 2nd, 3rd, and 4th quartiles (vs. 1st), respectively.

4.5.4 Sensitivity analyses

First, when we included only one randomly selected observation per subject from the final model there were no associations of UNGD activity metrics with BNP levels (**Table 4.6**). Using the same values, in this model there was an association (OR [95% CI]) of the 4th quartile (vs. 1st) of NDVI with high (vs. low) BNP, of 0.80 (0.66, 0.98). Second, we evaluated models weighted for the inverse probability of inclusion in the analysis in two ways, first using native inverse probability weights and second using truncated weights. These two analyses did not reveal substantive changes in associations (**Table 4.7, Model 1 and Figure 4.3a** [native] and **Model 2 and Figure 4.3b** [truncated]). Finally, we used linear regression to model *ln*-BNP on a continuous scale (**Table 4.8**) In this analysis, associations of the UNGD production metric with BNP were substantively the same as for the dichotomized analysis.

4.6 Discussion

To our knowledge, no prior studies have evaluated associations of such community and environmental factors as UNGD, greenness (using NDVI), and community social conditions leading to stress (using CSD) with BNP values in a large-scale, population-based epidemiologic study. BNP was considered to be an excellent candidate biomarker for this purpose because it is released by cardiac tissue under

stress due to ischemia, fibrosis, congestion, or cardiac remodeling [235] and can be measured in the blood. BNP is also associated with heart failure severity and survival [30, 123, 237]. Given our prior findings (**Chapter 3**) of associations of UNGD activity metrics with hospitalization among subjects with heart failure, we sought to evaluate whether these metrics were also associated with BNP. We hypothesized that UNGD metrics and CSD would be associated with greater blood BNP concentrations among heart failure patients, and that NDVI and distances to major and minor roads would be associated with lower blood BNP concentrations among heart failure patients.

We found consistent exposure-effect associations between quartiles of UNGD production activity and higher levels of BNP in both continuous and logistic models of 6501 BNP values among 3938 subjects with heart failure diagnoses from 2008 - 2015. We did not see any consistent associations between other UNGD metrics, NDVI, CSD, or distance to major and minor roads, and odds of BNP $\geq$ 400 pg/mL. We are confident that our models accounted for the various measured risk factors and comorbid conditions that are related to both heart failure and blood concentrations of BNP. These included sex, age, smoking status, relevant medication use (antihyperlipidemic and anticoagulant medications) at the time of the laboratory order, chronic kidney disease, BMI, and the Charlson index of morbidity. Additionally, we attempted to account for the duration of subjects' heart failure, by measuring the time (in days) from the first diagnosis code for heart failure in the EHR to the date of laboratory order.

Exposure-effect relations between UNGD production activity and higher BNP levels were present in sensitivity analyses that accounted for the inverse probability of being selected into this study and also in a linear model of *ln*-BNP. However, the association between UNGD production activity and BNP was not present in a sensitivity analysis that included only one randomly selected observation per person, utilizing 3938 laboratory orders as opposed to 6501 laboratory orders. Although this sensitivity

analysis was adequately powered, it is possible that we did not see an association with the production metric because we limited our analysis to only 61% of the available laboratory measurements, and the associations present in our main analysis were driven by the values for the remaining 39% of laboratory measurements. We speculate that these contrasting associations may also be because first BNP values were routinely obtained but subsequent measures were obtained due to concern of exacerbating heart failure. Because of this sensitivity analysis, the association present in the main analysis and in the inverse probability weighted analysis should be interpreted with caution. The contrasting ways that BNP is used and how heart failure hospitalization occurs also could explain, at least in part, the discordant results between this chapter and those in **Chapter 3**.

It is well-documented that UNGD is associated with a suite of environmental factors that can negatively impact community health, including air pollution (e.g., NO_x, $PM_{2.5}$, volatile organic compounds [VOCs], ozone), noise, traffic, and psychosocial stress [27, 132]. Of the four UNGD metrics that we calculated, the production metric is the only phase that did not end at a given well during our study period. Whereas the pad preparation, spud, and stimulation activity metrics reflect intermittent and limited-duration activities, the production metric cumulates over time as more and more wells are drilled, completed, enter production, and stay in production. The numerator for the production metric is the daily total volume of natural gas produced by the well, and once a well begins to produce natural gas, it continues to produce a nonzero volume on a daily basis. We speculate that a cumulating activity metric is more likely to evidence associations with a biomarker of heart failure than are the short-term activities at UNGD sites.

Another possible explanation for the contrasting associations of the UNGD activity metrics is that the UNGD production metric can also be an indicator for

compressor engine activity, which has been associated with increased noise levels [4]. Additionally, it is well documented that several air pollutants ($PM_{2.5}$, VOCs, NO_x, CO) are associated with the production phase of development [27, 233]. This might explain why we saw associations between the production metric and BNP but not the other UNGD metrics, as we observed for heart failure hospitalization in **Chapter 3**. However, since we have not been able to measure air pollutants, noise levels, or experiences of psychosocial stress directly, we cannot distinguish among these potential exposure and mechanistic pathways.

We suspect that we saw associations between multiple metrics (pad preparation, stimulation, and production) of UNGD activity and hospitalization for heart failure in **Chapter 3**, but only associations between UNGD production and odds of BNP $\geq$ 400 pg/mL in this analysis for several reasons. First, we believe that measurements of blood concentrations of BNP are more objective measures than hospitalization, which is dependent upon clinical decision making (i.e., a physician must make a determination to hospitalize a patient, a decision that could include such patient factors as frailty, social support, and access to the hospital) and heart failure management protocols that are designed to reduce health care costs by keeping patients out of the hospital. BNP can also be elevated among individuals with heart failure who are asymptomatic [237], and it has been associated with survival in individuals with and without heart failure [121]. Second, the different UNGD activity metrics reflect different exposure scenarios; pad preparation and stimulation activity metrics represent a substantial amount of truck traffic, whereas the production metric reflects compressor engine activity and the off-gassing of volatile organic compounds from produced natural gas. Third, although each UNGD activity metric was measured for a 30-day duration with a 1-day lag prior to the date of the BNP laboratory order, the production metric captures and cumulates ongoing UNGD activity, since production of natural gas at any individual well did not end during

our study period and total natural gas production in the state increased very dramatically during our study period. It is possible, then, that the association of UNGD production with BNP was the only one observed among the activity metrics because as it cumulated it was the only one to exceed a threshold of proximity to and intensity of sustained natural gas production. Fourth, it should be noted that the associations between the UNGD production metric were not present in a sensitivity analysis using only one observation per subject; thus the associations between UNGD production activity and BNP are driven primarily by individuals who have received multiple BNP measurements, and were presumably more severe in their heart failure, compared to individuals who only had one BNP measurement. Lastly, the laboratory measures in this analysis were obtained from both inpatient and outpatient settings, so this analysis reflects BNP levels from heart failure subjects who may not be experiencing an acute exacerbation, and who may be actively managing their heart failure through routine medical supervision and continued treatment.

Recent studies have called for more research into heart failure risk factors and outcomes because the public health burden of heart failure is increasing, and socioeconomic disparities in incidence have widened in recent years, suggesting that environmental factors could play a role in heart failure burden [210, 259]. Despite this, we have not been able to identify any large, population based studies of associations between environmental factors and BNP. Of the few studies that have assessed how environmental factors may be associated with BNP, generalizability of the results is limited by small sample sizes (i.e., n = 28 - 45) [124-126]. However, there is some evidence from these studies that air pollution exposure may impact BNP levels; in a study of both heart failure diagnosed subjects and healthy controls exposed to filtered and unfiltered diesel exhaust, the heart failure subjects had higher BNP levels and measures of endothelial dysfunction after exposure to unfiltered diesel exhaust, and that

filtered diesel exhaust reduced both measures of endothelial dysfunction and BNP [125]. Although this sample size was relatively small (n = 45), the findings support the biologic rationale for evaluating associations between environmental factors and BNP in a larger population. Similarly, a recent study in Colorado assessed cross-sectional associations of oil and gas activity with several other cardiac biomarkers in 249 healthy individuals and found that the highest levels of oil and gas activity and intensity were associated with higher blood pressure and slightly higher levels of IL-1β and α-TNF, two inflammatory markers [66]. This study considered oil production in addition to natural gas production and was also limited by a small sample size, but the findings support the plausible biologic pathway that implicates systemic inflammation in the progression of cardiac disease.

Our study limitations include the lack of information regarding individual socioeconomic status (e.g., household income), occupation, and reported well-being, all of which could impact an individual's heart failure prognosis [80, 143, 206]. A major limitation to this study is that we were not able to directly measure the air pollutants, noise levels, or psychosocial stress that we suspect are coming from the specific phases of UNGD. Therefore, it is not possible to disentangle psychosocial stress due to UNGD activity from the pollutants that UNGD activity is known to emit and which are known to have adverse cardiovascular impacts on exposed populations [24, 124, 129, 130, 260]. All three of these factors (air pollution, noise, psychosocial stress) are associated with systemic inflammation and endothelial dysfunction and thus could act on the same pathway leading to increased BNP. The lack of associations with NDVI could be due to our choice of peak greenness, as a feature of community context, rather than short-latency, time-varying measures of NDVI, for example, during the two weeks before the BNP measure. The lack of associations of CSD with BNP levels may be due to the fact that it is a measure of community context rather than a direct measure of individual-level

exposure to stress-causing events. Finally, future studies should evaluate network, rather than Euclidian, distances to health care facilties.

The study had several strengths, including the objective and independent assessment of environmental and community conditions and outcome; the large sample size; the general population representative sample; increasing covariate control; and systematic recording of data in the EHR for several years of observation. Furthermore, all of the BNP laboratory measurements were made at a central Geisinger laboratory facility, reducing the possibility of inter-laboratory variation in BNP measurements.

4.7 CONCLUSIONS

This is the first study to examine the clinical cardiac biomarker BNP in a large scale epidemiologic study of relevant environmental factors: UNGD activity, greenness, CSD, and distance to major and minor roads. In adjusted models, we observed exposure-effect relations of the UNGD production metric with high (vs. low) BNP values using a clinically-relevant cutoff [235]. The associations between greater UNGD activity and increased blood levels of BNP are biologically plausible given the functionality of BNP in response to exposure to air pollution, ventricular blood pressure, and subsequent acute cardiac myocyte stress [107, 125, 235, 239]. This has public health relevance because higher BNP levels have been associated with poor prognosis and mortality among individuals with and without heart failure [121]. Lastly, these findings are consistent with previous epidemiologic studies of UNGD activity phases and objectively measured health outcomes, adding to a growing body of literature supporting the negative impacts of UNGD on population health [13, 15, 17, 18, 65].

4.8 REFERENCES

1. Magnussen, C. and S. Blankenberg, *Biomarkers for heart failure: small molecules with high clinical relevance.* Journal of Internal Medicine, 2018. **283**(6): p. 530-543.
2. Gaggin, H. and J. Januzzi. *Cardiac biomarkers and heart failure*. Latest in Cardiology 2015 [cited 2018 7/26/18]; Available from: https://www.acc.org/latest-in-cardiology/articles/2015/02/09/13/00/cardiac-biomarkers-and-heart-failure.
3. Maisel, A.S., J.M. Duran, and N. Wettersten, *Natriuretic Peptides in Heart Failure: Atrial and B-type Natriuretic Peptides.* Heart Failure Clinics, 2018. **14**(1): p. 13-25.
4. Wang, T.J., et al., *Plasma Natriuretic Peptide Levels and the Risk of Cardiovascular Events and Death.* New England Journal of Medicine, 2004. **350**(7): p. 655-663.
5. York, M.K., et al., *B-Type Natriuretic Peptide Levels and Mortality in Patients With and Without Heart Failure.* Journal of the American College of Cardiology, 2018. **71**(19): p. 2079-2088.
6. Kotecha, D.A.-O.h.o.o., et al., *B-type natriuretic peptide trumps other prognostic markers in patients assessed for coronary disease.* (1741-7015 (Electronic)).
7. Potter, L.R., *Natriuretic peptide metabolism, clearance and degradation.* The FEBS journal, 2011. **278**(11): p. 1808-1817.
8. Sarzani, R.A.-O.h.o.o., et al., *Cardiac Natriuretic Peptides, Hypertension and Cardiovascular Risk.* (1179-1985 (Electronic)).
9. Jordan, J., et al., *Natriuretic Peptides in Cardiovascular and Metabolic Crosstalk: Implications for Hypertension Management.* (1524-4563 (Electronic)).
10. Rajagopalan, S., S.G. Al-Kindi, and R.D. Brook, *Air Pollution and Cardiovascular Disease: JACC State-of-the-Art Review.* (1558-3597 (Electronic)).
11. Shah, A.S.V., et al., *Global association of air pollution and heart failure: a systematic review and meta-analysis.* The Lancet. **382**(9897): p. 1039-1048.
12. Li, W., et al., *Short-Term Exposure to Ambient Air Pollution and Biomarkers of Systemic Inflammation.* Arteriosclerosis, Thrombosis, and Vascular Biology, 2017. **37**(9): p. 1793-1800.
13. Pope, C.A., et al., *Exposure to Fine Particulate Air Pollution Is Associated With Endothelial Injury and Systemic Inflammation.* Circulation Research, 2016. **119**(11): p. 1204-1214.
14. Krishnan, R.M., et al., *Vascular Responses to Long- and Short-Term Exposure to Fine Particulate Matter: MESA Air (Multi-Ethnic Study of Atherosclerosis and Air Pollution).* Journal of the American College of Cardiology, 2012. **60**(21): p. 2158-2166.
15. Sanidas, E., et al., *Air pollution and arterial hypertension. A new risk factor is in the air.* (1878-7436 (Electronic)).
16. Giorgini, P., et al., *Air Pollution Exposure and Blood Pressure: An Updated Review of the Literature.* (1873-4286 (Electronic)).
17. Ushakov Av Fau - Ivanchenko, V.S., V.S. Ivanchenko, and A.A. Gagarina, *Psychological Stress in Pathogenesis of Essential Hypertension.* (1875-6506 (Electronic)).
18. Schmidt, F., et al., *Nighttime aircraft noise impairs endothelial function and increases blood pressure in patients with or at high risk for coronary artery disease.* (1861-0692 (Electronic)).

19. Munzel, T., et al., *Environmental stressors and cardio-metabolic disease: part I-epidemiologic evidence supporting a role for noise and air pollution and effects of mitigation strategies.* (1522-9645 (Electronic)).
20. Cai, Y., et al., *Long-term exposure to road traffic noise, ambient air pollution, and cardiovascular risk factors in the HUNT and lifelines cohorts.* (1522-9645 (Electronic)).
21. Wellenius, G.A., et al., *Effects of ambient air pollution on functional status in patients with chronic congestive heart failure: a repeated-measures study.* (1476-069X (Electronic)).
22. Emiroglu, Y., et al., *BNP levels in patients with long-term exposure to biomass fuel and its relation to right ventricular function.* (1522-9629 (Electronic)).
23. Vieira, J.L., et al., *Respiratory Filter Reduces the Cardiovascular Effects Associated With Diesel Exhaust Exposure: A Randomized, Prospective, Double-Blind, Controlled Study of Heart Failure: The FILTER-HF Trial.* (2213-1787 (Electronic)).
24. Zilz, C., et al., *Mental health, serum biomarkers and survival in severe COPD: a pilot study.* Multidisciplinary respiratory medicine, 2016. **11**: p. 3-3.
25. Wawrzyniak, A.J., et al., *High Concordance Between Mental Stress-Induced and Adenosine-Induced Myocardial Ischemia Assessed Using SPECT in Heart Failure Patients: Hemodynamic and Biomarker Correlates.* Journal of nuclear medicine : official publication, Society of Nuclear Medicine, 2015. **56**(10): p. 1527-1533.
26. Pujades-Rodriguez, M., et al., *Socioeconomic Deprivation and the Incidence of 12 Cardiovascular Diseases in 1.9 Million Women and Men: Implications for Risk Prediction and Prevention.* PLoS ONE, 2014. **9**(8): p. e104671.
27. Barrington, W.E., et al., *Neighborhood socioeconomic deprivation, perceived neighborhood factors, and cortisol responses to induced stress among healthy adults.* Health Place, 2014. **27**: p. 120-6.
28. Yeager, R., et al., *Association Between Residential Greenness and Cardiovascular Disease Risk.* Journal of the American Heart Association, 2018. **7**(24): p. e009117.
29. Wang, K., et al., *Relationship of Neighborhood Greenness to Heart Disease in 249 405 US Medicare Beneficiaries.* Journal of the American Heart Association, 2019. **8**(6): p. e010258-e010258.
30. Brown, S.C., et al., *Neighborhood Greenness and Chronic Health Conditions in Medicare Beneficiaries.* American Journal of Preventive Medicine, 2016. **51**(1): p. 78-89.
31. van Veldhuisen, D.J., et al., *B-Type Natriuretic Peptide and Prognosis in Heart Failure Patients With Preserved and Reduced Ejection Fraction.* Journal of the American College of Cardiology, 2013. **61**(14): p. 1498-1506.
32. Medi-Span, *Master Drug Data Base Documentation Manual*. 2007, Indianapolis, Indiana.
33. Casey, J.A., et al., *Unconventional Natural Gas Development and Birth Outcomes in Pennsylvania, USA.* Epidemiology, 2015.
34. Casey, J.A., et al., *Measures of SES for Electronic Health Record-based Research.* American Journal of Preventive Medicine, 2018. **54**(3): p. 430-439.
35. Koehler, K., et al., *Exposure Assessment Using Secondary Data Sources in Unconventional Natural Gas Development and Health Studies.* Environmental Science & Technology, 2018. **52**(10): p. 6061-6069.
36. *American Community Survey*. 2010, United States Census Bureau.

37. Nau, C., et al., *Community socioeconomic deprivation and obesity trajectories in children using electronic health records.* Obesity, 2014. **23**(1): p. 207-212.
38. Rasmussen, S.G., et al., *Association Between Unconventional Natural Gas Development in the Marcellus Shale and Asthma Exacerbations. .* JAMA Internal Medicine, 2016(2168-6114).
39. Casey, J.A., et al., *Greenness and Birth Outcomes in a Range of Pennsylvania Communities.* International Journal of Environmental Research and Public Health, 2016. **13**(3).
40. Wu, A.H.B., *Serial testing of B-type natriuretic peptide and NTpro-BNP for monitoring therapy of heart failure: The role of biologic variation in the interpretation of results.* American Heart Journal, 2006. **152**(5): p. 828-834.
41. Oakes, J.M., *The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology.* (0277-9536 (Print)).
42. Diez Roux, A.V., *Estimating neighborhood health effects: the challenges of causal inference in a complex world.* (0277-9536 (Print)).
43. Hubbard, A.E., et al., *To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health.* (1531-5487 (Electronic)).
44. Pan, W.E.I., *Goodness-of-fit Tests for GEE with Correlated Binary Data.* Scandinavian Journal of Statistics, 2002. **29**(1): p. 101-110.
45. Gregg, L.P., et al., *Effect Modification of Chronic Kidney Disease on the Association of Circulating and Imaging Cardiac Biomarkers With Outcomes. LID - e005235 [pii] LID - 10.1161/JAHA.116.005235 [doi].* (2047-9980 (Electronic)).
46. Vodovar, N. and D. Logeart, *Similar BNP and Mortality Association in Patients With and Without Heart Failure.* Journal of the American College of Cardiology, 2018. **71**(19): p. 2089.
47. Adgate, J.L., B.D. Goldstein, and L.M. McKenzie, *Potential Public Health Hazards, Exposures and Health Effects from Unconventional Natural Gas Development.* Environmental Science & Technology, 2014. **48**(15): p. 8307-8320.
48. Boyle, M.D., et al., *A pilot study to assess residential noise exposure near natural gas compressor stations.* PloS one, 2017. **12**(4): p. e0174310-e0174310.
49. Srebotnjak, T. and M. Rotkin-Ellman, *Fracking Fumes: Air pollution from hydraulic fracturing threatens public health and communities*. 2014, Natural Resources Defense Council;.
50. York, M.K., et al., *B-Type Natriuretic Peptide Levels and Mortality in Patients With and Without Heart Failure.* (1558-3597 (Electronic)).
51. Zannad, F., *Rising incidence of heart failure demands action.* The Lancet, 2018. **391**(10120): p. 518-519.
52. Conrad, N., et al., *Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals.* The Lancet, 2018. **391**(10120): p. 572-580.
53. McKenzie, L.M., et al., *Relationships between indicators of cardiovascular disease and intensity of oil and natural gas activity in Northeastern Colorado.* Environmental Research, 2019. **170**: p. 56-64.
54. Ogilvie, R.P., et al., *Psychosocial Factors and Risk of Incident Heart Failure: The Multi-Ethnic Study of Atherosclerosis.* Circ Heart Fail, 2016. **9**(1): p. e002243.
55. Nieminen, M.S., et al., *The patient perspective: Quality of life in advanced heart failure with frequent hospitalisations.* International Journal of Cardiology, 2015. **191**: p. 256-264.

56. Friedmann, E., et al., *Relationship of depression, anxiety, and social isolation to chronic heart failure outpatient mortality.* American Heart Journal, 2006. **152**(5): p. 940.e1-940.e8.
57. Rodopoulou, S., et al., *Air pollution and cardiovascular and respiratory emergency visits in Central Arkansas: A time-series analysis.* (1879-1026 (Electronic)).
58. Swarthout, R.F., et al., *Impact of Marcellus Shale Natural Gas Development in Southwest Pennsylvania on Volatile Organic Compound Emissions and Regional Air Quality.* Environmental Science & Technology, 2015. **49**(5): p. 3175-3184.
59. Wellenius, G.A., et al., *Particulate Air Pollution and the Rate of Hospitalization for Congestive Heart Failure among Medicare Beneficiaries in Pittsburgh, Pennsylvania.* American Journal of Epidemiology, 2005. **161**(11): p. 1030-1036.
60. Roy, A., P. Adams, and A. Robinson, *Air pollutant emissions from the development, production and processing of Marcellus Shale Gas.* Journal of the Air and Waste Management Association, 2014. **64**(1): p. 19-37.
61. Tustin, A.W., et al., *Associations between Unconventional Natural Gas Development and Nasal and Sinus, Migraine Headache, and Fatigue Symptoms in Pennsylvania.* Environ Health Perspect, 2016.
62. Walker, W.K., K.A. Marshall, and E. Symanski, *Drilling and Production Activity Related to Unconventional Gas Development and Severity of Preterm Birth.* Environ Health Perspect. , 2018. **126**(3): p. 037006.
63. Casey, J.A., et al., *Associations of unconventional natural gas development with depression symptoms and disordered sleep in Pennsylvania.* Scientific Reports, 2018. **8**(1): p. 11375.

Figure 4.1a. Histogram of unadjusted B-type natriuretic peptide (BNP) values (n = 6501)

Figure 4.1b. Histogram of unadjusted B-type natriuretic peptide (BNP) values after *ln*-transformation (*ln*-BNP) (n = 6501)

Figure 4.2a. Greenness using normalized difference vegetation index (NDVI) quartiles assigned to residential address of subjects' with B-type natriuretic peptide (BNP) measures in study area in Pennsylvania

Figure 4.2b. Unconventional natural gas development (UNGD) pad activity metric quartiles assigned to residential address of subjects' with B-type natriuretic peptide (BNP) measures in study area in Pennsylvania.

Figure 4.2c. Unconventional natural gas development (UNGD) spud activity metric quartiles assigned to residential address of subjects' with B-type natriuretic peptide (BNP) measures in study area in Pennsylvania.
Figure 4.2d. Unconventional natural gas development (UNGD) stimulation activity metric quartiles assigned to residential address of subjects' with B-type natriuretic peptide (BNP) measures in study area in Pennsylvania.
Figure 4.2e. Unconventional natural gas development (UNGD) production activity metric quartiles assigned to residential address of subjects' with B-type natriuretic peptide (BNP) measures in study area in Pennsylvania.

Figure 4.3a. Histogram of inverse probability weights to account for selection into the study. The model to generate these weights adjusted for: smoking status, time of contact with the electronic health records (EHRs), Charlson index, age, diagnosis of myocardial infarction, coronary artery disease, valve disease, chronic kidney disease, and anticoagulant medication. Subjects with very high weights in the right tail of this distribution had a very low probability of being selected into the BNP study according to this model (i.e., these subjects did not have many comorbidities, were generally younger, and had less time of observation in the EHR than those who had a higher probability of being selected into the study). Thus, the inverse probability weights for these subjects are very high.

Figure 4.3b. Histogram of inverse probability weights to account for selection, truncated at the 99th percentile of weights for all 3938 subjects (maximum inverse probability weight = 9.5).

Table 4.1. Number of B-type natriuretic peptide (BNP) values per subject

BNP values in analysis per person	Subjects*, n (%)
1	2441 (62.0)
2	858 (21.8)
3	340 (8.6)
4	171 (4.3)
5	128 (3.3)
Total	**3938**

*Number of unique subjects by number of BNP measure

Table 4.2. Descriptive statistics of study population by first B-type natriuretic peptide (BNP) laboratory order, by setting, n = 3938

	Laboratory Order Setting		
	Outpatient	Inpatient	
Variable	n = 1703	n = 2235	p-value*
Age, mean (SD)	72.2 (11.6)	70.7 (12.3)	< 0.001
Charlson index, mean (SD)	8.7 (2.7)	8.9 (3.1)	0.01
Sex, n (%)			
Female	815 (47.9)	1026 (45.9)	
Male	888 (52.1)	1209 (54.1)	0.2
Race/ethnicity, n (%)			
White	1664 (97.7)	2170 (97.1)	
Non-white	39 (2.3)	65 (2.9)	0.2
Community type, n (%)			
Borough	544 (31.4)	743 (33.2)	
Township	1004 (59.0)	1192 (53.3)	
Census tract (city)	155 (9.1)	300 (13.4)	< 0.001
Receipt of Medical Assistance, n (%)			
Prior to laboratory date	149 (8.8)	259 (11.6)	0.004
Body mass index (BMI) at event, kg/m^2, mean (SD)	32.3 (7.0)	32.0 (7.7)	0.3
Smoking status, n (%)			
Ever	996 (58.5)	1355 (60.6)	
Never	707 (41.5)	880 (39.4)	0.2
Antihyperlipidemic medication, n (%)			
At laboratory date	895 (55.6)	951 (42.6)	< 0.001
Anticoagulant medication, n (%)			
At laboratory date	406 (23.8)	446 (20.0)	0.003
Antihypertensive medication, n (%)	838 (49.2)	847 (37.9)	< 0.001

At laboratory date			
Diagnosis of chronic kidney disease, n (%)			
Prior to laboratory date	835 (49.0)	846 (37.9)	< 0.001
Diagnosis of myocardial infarction, n (%)			
Prior to laboratory date	135 (7.9)	225 (10.1)	0.02
Diagnosis of valve disease, n (%)			
Prior to laboratory date	319 (18.7)	432 (19.3)	0.6
Diagnosis of chronic obstructive pulmonary disease, n (%)			
Prior to laboratory date	322 (18.9)	527 (23.6)	< 0.001
Diagnosis of hypertension, n (%)			
Prior to laboratory date	1005 (59.0)	1566 (70.1)	< 0.001
Diagnosis of diabetes, n (%)			
Prior to laboratory date	592 (34.8)	985 (44.1)	< 0.001
Distance to nearest hospital or clinic, mean (SD), meters	6890.6 (7866.3)	6213.4 (8104.8)	0.009
Distance to nearest major road, mean (SD), meters	2502.8 (3666.8)	2743.9 (4412.5)	0.07
Distance to nearest minor road, mean (SD), meters	1795.2 (2646.6)	1324.7 (2014.9)	< 0.001
Pad preparation metric, n (%),1/m^2 **			
Q1 (1.22 X 10^{-10}, 1.76 X 10^{-9})	577 (33.9)	668 (29.9)	
Q2 (1.76 X 10^{-9}, 3.61 X 10^{-9})	476 (28.0)	493 (22.1)	
Q3 (3.61 X 10^{-9}, 6.04 X 10^{-9})	336 (19.7)	528 (23.6)	
Q4 (6.06 X 10^{-9}, 3.75 X 10^{-6})	314 (18.4)	546 (24.4)	< 0.001
Spud metric, n (%),1/m^2 **			
Q1 (4.9 X 10^{-12}, 4.13 X 10^{-11})	583 (34.2)	751 (33.6)	
Q2 (4.2 X 10^{-11}, 1.17 X 10^{-10})	383 (22.5)	558 (25.0)	

Q3 (1.17 X 10^{-10}, 1.59 X 10^{-10})	325 (19.1)	530 (23.7)	
Q4 (1.59 X 10^{-10}, 4.15 X 10^{-10})	412 (24.2)	396 (17.7)	< 0.001
Stimulation metric, n (%), m/m^2 **			
Q1 (0, 2.49 X 10^{-6})	581 (34.1)	702 (31.4)	
Q2 (2.50 X 10^{-6}, 1.24 X 10^{-5})	445 (26.1)	529 (23.7)	
Q3 (1.24 X 10^{-5}, 2.68 X 10^{-5})	373 (21.9)	463 (20.7)	
Q4 (2.68 X 10^{-5}, 4.04 X 10^{-3})	304 (17.9)	541 (24.2)	< 0.001
Production metric, n (%), m^3/m^2 **			
Q1 (1.03 X 10^{-7}, 9.64 X 10^{-6})	608 (35.7)	732 (32.8)	
Q2 (9.69 X 10^{-6}, 8.95 X 10^{-5})	409 (24.0)	521 (23.3)	
Q3 (8.97 X 10^{-5}, 3.07 X 10^{-4})	364 (21.4)	460 (20.6)	
Q4 (3.07 X 10^{-4}, 0.102)	322 (18.9)	522 (23.4)	0.008
NDVI***, n (%), unitless			
Q1 (0.17, 0.58)	359 (21.1)	620 (27.7)	
Q2 (0.58, 0.68)	412 (24.2)	526 (23.5)	
Q3 (0.68, 0.77)	457 (26.8)	548 (24.5)	
Q4 (0.77, 0.92)	475 (27.9)	541 (24.2)	< 0.001
CSD†, n (%), SD units			
Q1 (-14.6, -8.3)	418 (24.5)	577 (25.8)	
Q2 (-8.3, - 5.9)	462 (27.1)	527 (23.6)	
Q3 (-5.9, -3.1)	397 (23.3)	566 (25.3)	
Q4 (-3.0, 18.2)	426 (25.0)	565 (25.3)	0.07

*p-value obtained from one-way analysis of variance (ANOVA) F-test for continuous variables, chi^2 test statistic for categorical variables

** Ranges for quartiles of UNGD activity metrics are listed next to each quartile

***Normalized difference vegetation index (NDVI), a measure of greenness, was measured by a 250 m x 250 m grid surrounding a subject's residential address for the period of maximum annual greenness in the year of each laboratory measurement

†Community socioeconomic deprivation (CSD), measured from the 2010 US Census American Community Survey at each subject's designated place.

Table 4.3. Unadjusted bivariate associations of selected categorical variables with B-type natriuretic peptide (BNP) values on a continuous scale

Variable	Number (%) out of 6501 values	Mean BNP value (pg/mL), SD	p-value*
Sex			
Female	3014 (46.4)	450.6 (590.6)	
Male	3487 (53.6)	503.3 (674.6)	0.1
Race/ethnicity			
White	6346 (97.6)	479.0 (636.5)	
Nonwhite	155 (2.4)	470.0 (682.3)	0.02
Community type			
Borough	2114 (32.5)	477.1 (645.6)	
Township	3646 (56.1)	468.3 (616.0)	
Census tract (city)	741 (11.4)	535.6 (712.0)	0.4
Laboratory setting			
Outpatient	3072 (47.3)	330.6 (439.7)	
Inpatient	3429 (52.8)	611.6 (748.5)	< 0.001
Receipt of Medical Assistance			
Ever prior to laboratory date	761 (11.7)	423.7 (636.5)	
Never prior to laboratory date	5740 (88.3)	486.1 (637.3)	< 0.001
Hypertension diagnosis**			
Yes	4336 (66.7)	496.8 (656.3)	
No	2165 (33.3)	442.9 (596.8)	< 0.001
Type 2 diabetes diagnosis**			
Yes	2762 (42.5)	489.8 (645.6)	
No	3739 (57.5)	470.7 (631.5)	0.01
Myocardial infarction diagnosis**			
Yes	687 (10.6)	618.2 (707.6)	
No	5814 (89.4)	462.4 (626.8)	< 0.001

Chronic obstructive pulmonary disease diagnosis**			
Yes	1653 (25.4)	437.5 (603.8)	
No	4848 (74.6)	492.9 (648.1)	0.04
Valve disorder diagnosis**			
Yes	1490 (22.9)	557.4 (678.3)	
No	5011 (77.1)	455.5 (623.1)	< 0.001
Coronary artery disease diagnosis**			
Yes	1625 (25.0)	523.5 (675.8)	
No	4876 (75.0)	464.0 (623.6)	0.002
Smoking status			
Ever	4029 (62.0)	464.6 (632.1)	
Never	2472 (38.0)	502.1 (645.8)	< 0.001
Antihyperlipidemic medication, at time of laboratory date			
Yes	3314 (51.0)	460.5 (622.6)	
No	3187 (49.0)	497.9 (652.3)	0.03
Anticoagulant medication, at time of laboratory date			
Yes	1703 (26.2)	488.4 (593.3)	
No	4798 (73.8)	475.4 (652.6)	< 0.001
Antihypertensive medication, at time of laboratory date			
Yes	3161 (48.6)	447.5 (599.7)	
No	3340 (51.4)	508.5 (670.1)	0.002
Diagnosis of chronic kidney disease**			
Yes	3214 (49.4)	542.0 (696.9)	
No	3287 (50.6)	417.1 (566.9)	< 0.001

Pad preparation metric, n (%),1/m^2***			
Q1 (1.22 X 10^{-10}, 1.76 X 10^{-9})	1626 (25.0)	498.5 (700.6)	
Q2 (1.76 X 10^{-9}, 3.61 X 10^{-9})	1625 (25.0)	455.4 (616.4)	
Q3 (3.61 X 10^{-9}, 6.04 X 10^{-9})	1626 (25.0)	499.4 (613.8)	
Q4 (6.06 X 10^{-9}, 3.75 X 10^{-6})	1624 (25.0)	462.0 (614.3)	0.005
Spud metric, n (%),1/m^2 ***			
Q1 (4.9 X 10^{-12}, 4.13 X 10^{-11})	1626 (25.0)	498.1 (692.6)	
Q2 (4.2 X 10^{-11}, 1.17 X 10^{-10})	1625 (25.0)	510.2 (667.7)	
Q3 (1.17 X 10^{-10}, 1.59 X 10^{-10})	1625 (25.0)	463.4 (613.1)	
Q4 (1.59 X 10^{-10}, 4.15 X 10^{-10})	1625 (25.0)	443.7 (567.6)	0.02
Stimulation metric, n (%), m/m^2 ***			
Q1 (0, 2.49 X 10^{-6})	1626 (25.0)	488.7 (656.9)	
Q2 (2.50 X 10^{-6}, 1.24 X 10^{-5})	1625 (25.0)	503.4 (687.9)	
Q3 (1.24 X 10^{-5}, 2.68 X 10^{-5})	1626 (25.0)	450.5 (595.8)	
Q4 (2.68 X 10^{-5}, 4.04 X 10^{-3})	1624 (25.0)	472.7 (604.4)	0.2
Production metric, n (%), m^3/m^2 ***			
Q1 (1.03 X 10^{-7}, 9.64 X 10^{-6})	1626 (25.0)	484.9 (675.1)	
Q2 (9.69 X 10^{-6}, 8.95 X 10^{-5})	1625 (25.0)	507.8 (660.1)	
Q3 (8.97 X 10^{-5}, 3.07 X 10^{-4})	1626 (25.0)	458.4 (618.4)	
Q4 (3.07 X 10^{-4}, 0.102)	1624 (25.0)	464.1 (592.5)	0.002
NDVI†, n (%), unitless			
Q1 (0.17, 0.58)	1626 (25.0)	511.8 (698.0)	
Q2 (0.58, 0.68)	1626 (25.0)	467.5 (606.5)	
Q3 (0.68, 0.77)	1624 (25.0)	482.1 (613.0)	
Q4 (0.77, 0.92)	1625 (25.0)	453.7 (627.7)	0.1

CSD††, n (%), SD units			
Q1 (-14.6, -8.3)	1637 (25.2)	481.1 (639.6)	
Q2 (-8.3, - 5.9)	1647 (25.3)	465.8 (638.7)	
Q3 (-5.9, -3.1)	1592 (24.5)	465.4 (619.1)	
Q4 (-3.0, 18.2)	1625 (25.0)	503.0 (651.8)	0.8

*p-value obtained from unadjusted generalized estimating equations with *ln*-BNP as the outcome

** Diagnoses were classified as yes if subject had at least two diagnosis codes for the condition prior to the laboratory date

*** Ranges for quartiles of UNGD activity metrics are listed next to each quartile

†Normalized difference vegetation index (NDVI), a measure of greenness, was measured by a 250 m x 250 m grid surrounding a subject's residential address for the period of maximum annual greenness in the year of each laboratory measurement

††Community socioeconomic deprivation (CSD), measured from the 2010 US Census American Community Survey at each subject's designated place.

Table 4.4. Descriptive statistics of study population at the time of the first BNP laboratory order, by diagnostic cutoff of 400 pg/mL, n = 3838

	Laboratory value		
	< 400 pg/mL	≥ 400 pg/mL	
Variable	n = 2459	n = 1479	p-value*
Age, mean (SD)	71.0 (12.2)	73.7 (11.6)	< 0.001
Charlson index, mean (SD)	8.6 (3.2)	9.7 (3.3)	< 0.001
Sex, n (%)			
Female	1182 (48.1)	659 (44.6)	
Male	1277 (51.9)	820 (55.4)	0.03
Race/ethnicity, n (%)			
White	2390 (97.2)	1444 (97.6)	
Non-white	69 (2.8)	35 (2.4)	0.4
Community type, n (%)			
Borough	781 (31.8)	506 (34.2)	
Township	1419 (57.7)	777 (52.5)	
Census tract (city)	259 (10.5)	196 (13.3)	0.002
Receipt of Medical Assistance, n (%)			
Prior to laboratory date	282 (11.5)	126 (8.5)	0.003
Body mass index (BMI) at event, kg/m^2, mean (SD)	33.4 (8.4)	29.5 (7.7)	< 0.001
Smoking status, n (%)			
Ever	1555 (63.2)	875 (59.2)	
Never	904 (36.8)	604 (40.8)	0.01
Antihyperlipidemic medication, n (%)			
At laboratory date	1369 (55.7)	744 (50.3)	0.001
Anticoagulant medication, n (%)			
At laboratory date	673 (27.4)	440 (29.8)	0.1
Antihypertensive medication, n (%)			<
At laboratory date	1288 (52.4)	676 (45.7)	0.001
Diagnosis of chronic kidney disease, n (%)			<
Prior to laboratory date	1132 (46.0)	817 (55.2)	0.001
Diagnosis of myocardial infarction, n (%)			<
Prior to laboratory date	223 (9.1)	214 (14.5)	0.001
Diagnosis of valve disease, n (%)			<
Prior to laboratory date	453 (18.4)	390 (26.4)	0.001

Diagnosis of chronic obstructive pulmonary disease, n (%)			
Prior to laboratory date	627 (25.5)	350 (23.7)	0.2
Diagnosis of hypertension, n (%)			
Prior to laboratory date	1583 (64.4)	1014 (68.6)	0.007
Diagnosis of diabetes, n (%)			
Prior to laboratory date	988 (40.2)	623 (42.1)	0.2
Distance to nearest hospital or clinic, mean (SD), meters	6574 (7834)	6395 (8292)	0.5
Distance to nearest major road, mean (SD), meters	2650 (4008)	2622 (4270)	0.8
Distance to nearest minor road, mean (SD), meters	1615 (2448)	1384 (2085)	0.003
Pad preparation metric, n (%),$1/m^2$ **			
Q1 (1.22 X 10^{-10}, 1.76 X 10^{-9})	477 (19.4)	297 (20.1)	
Q2 (1.76 X 10^{-9}, 3.61 X 10^{-9})	650 (26.4)	354 (23.9)	
Q3 (3.61 X 10^{-9}, 6.04 X 10^{-9})	647 (26.3)	431 (29.1)	
Q4 (6.06 X 10^{-9}, 3.75 X 10^{-6})	685 (27.9)	397 (26.8)	0.1
Spud metric, n (%),$1/m^2$ **			
Q1 (4.9 X 10^{-12}, 4.13 X 10^{-11})			
Q2 (4.2 X 10^{-11}, 1.17 X 10^{-10})			
Q3 (1.17 X 10^{-10}, 1.59 X 10^{-10})	435 (17.7)	277 (18.7)	
Q4 (1.59 X 10^{-10}, 4.15 X 10^{-10})	544 (22.1)	363 (24.5)	
	708 (28.8)	411 (27.8)	
	772 (31.4)	428 (28.9)	0.2
Stimulation metric, n (%), m/m^2 **			
Q1 (0, 2.49 X 10^{-6})	437 (17.8)	276 (18.7)	
Q2 (2.50 X 10^{-6}, 1.24 X 10^{-5})	517 (21.0)	331 (22.4)	
Q3 (1.24 X 10^{-5}, 2.68 X 10^{-5})	705 (28.7)	395 (26.7)	
Q4 (2.68 X 10^{-5}, 4.04 X 10^{-3})	800 (32.5)	477 (32.3)	
			0.5
Production metric, n (%), m^3/m^2 **			
Q1 (1.03 X 10^{-7}, 9.64 X 10^{-6})	442 (18.0)	275 (18.6)	
Q2 (9.69 X 10^{-6}, 8.95 X 10^{-5})	494 (20.1)	317 (21.4)	
Q3 (8.97 X 10^{-5}, 3.07 X 10^{-4})	679 (27.6)	397 (26.8)	
Q4 (3.07 X 10^{-4}, 0.102)	844 (34.3)	490 (33.1)	
			0.7
NDVI***, n (%), unitless			
Q1 (0.17, 0.58)	578 (23.5)	397 (26.8)	
Q2 (0.58, 0.68)	624 (25.4)	363 (24.5)	
Q3 (0.68, 0.77)	612 (24.9)	369 (25.0)	
Q4 (0.77, 0.92)	645 (26.2)	250 (23.7)	0.08
CSD†, n (%), SD units	614 (25.0)	381 (25.8)	
Q1 (-14.6, -8.3)	628 (25.5)	361 (24.4)	0.4

Q2 (-8.3, - 5.9)	616 (25.1)	347 (23.5)
Q3 (-5.9, -3.1)	601 (24.4)	390 (26.4)
Q4 (-3.0, 18.2)		

*p-value obtained from one-way analysis of variance (ANOVA) F-test for continuous variables, chi^2 test statistic for categorical variables

** Ranges for quartiles of UNGD activity metrics are listed next to each quartile

***Normalized difference vegetation index (NDVI), a measure of greenness, was measured by a 250 m x 250 m grid surrounding a subject's residential address for the period of maximum annual greenness in the year of each laboratory measurement

†Community socioeconomic deprivation (CSD), measured from the 2010 US Census American Community Survey at each subject's designated place.

Figure 4.1a

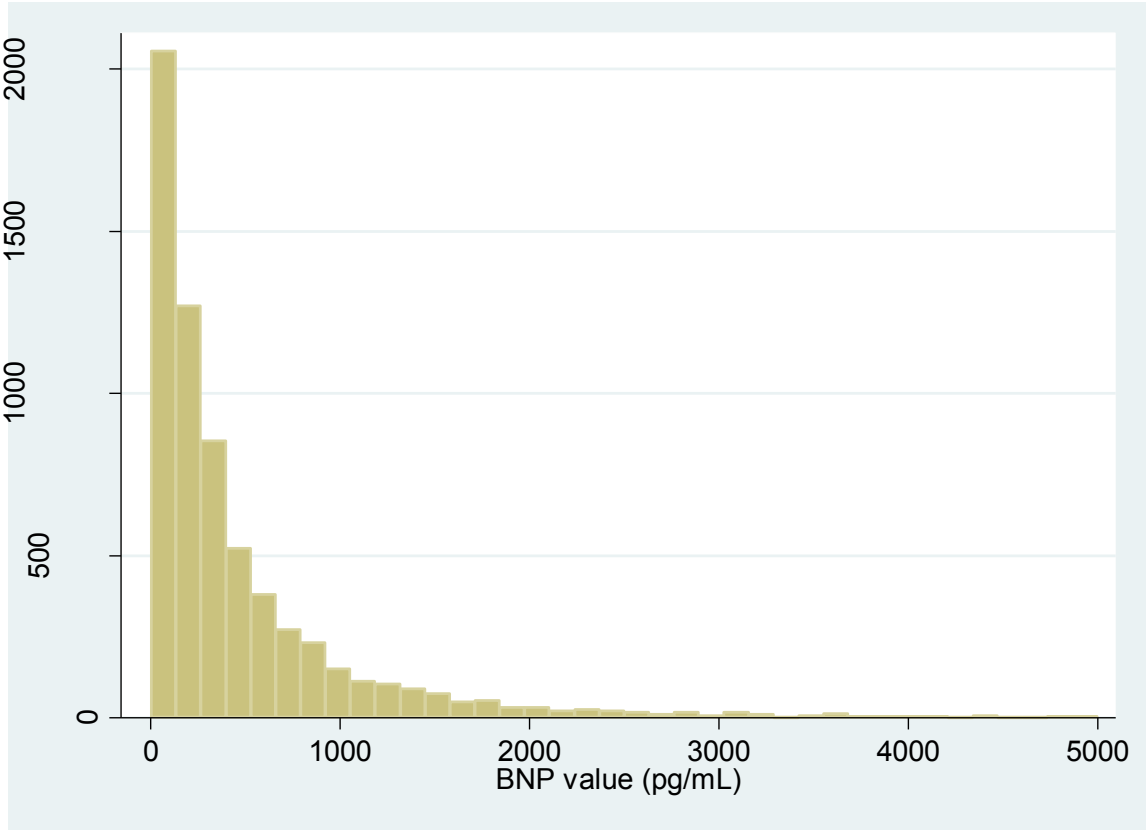


Figure 4.1b

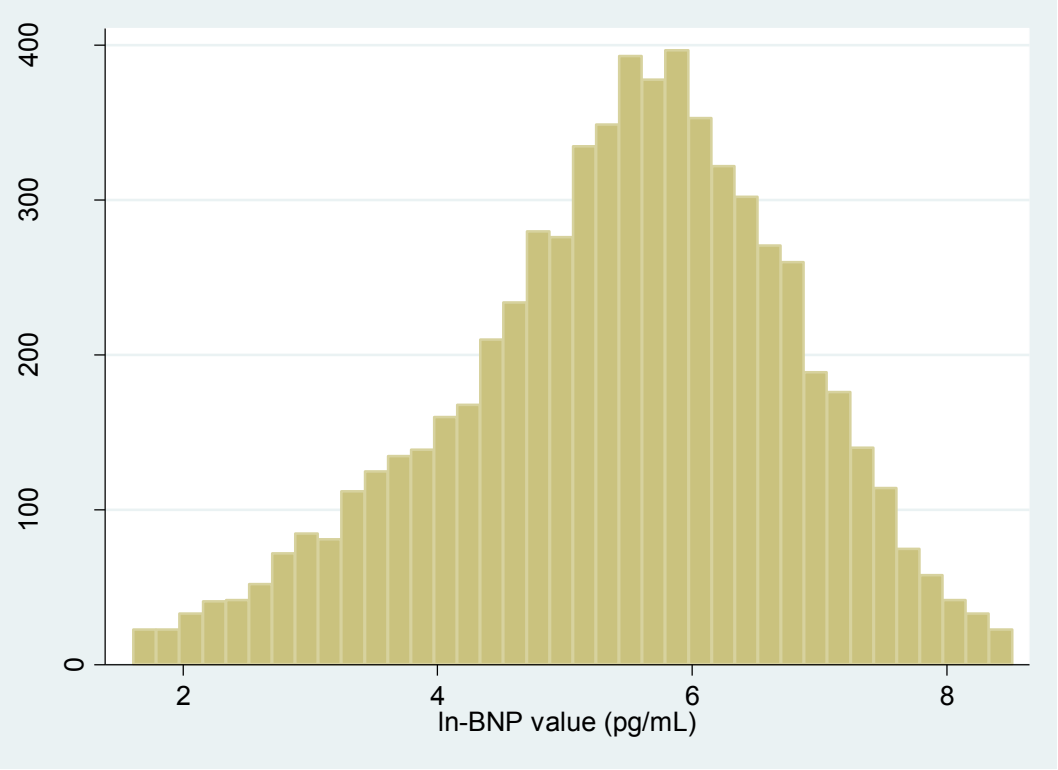


Figure 4.2a

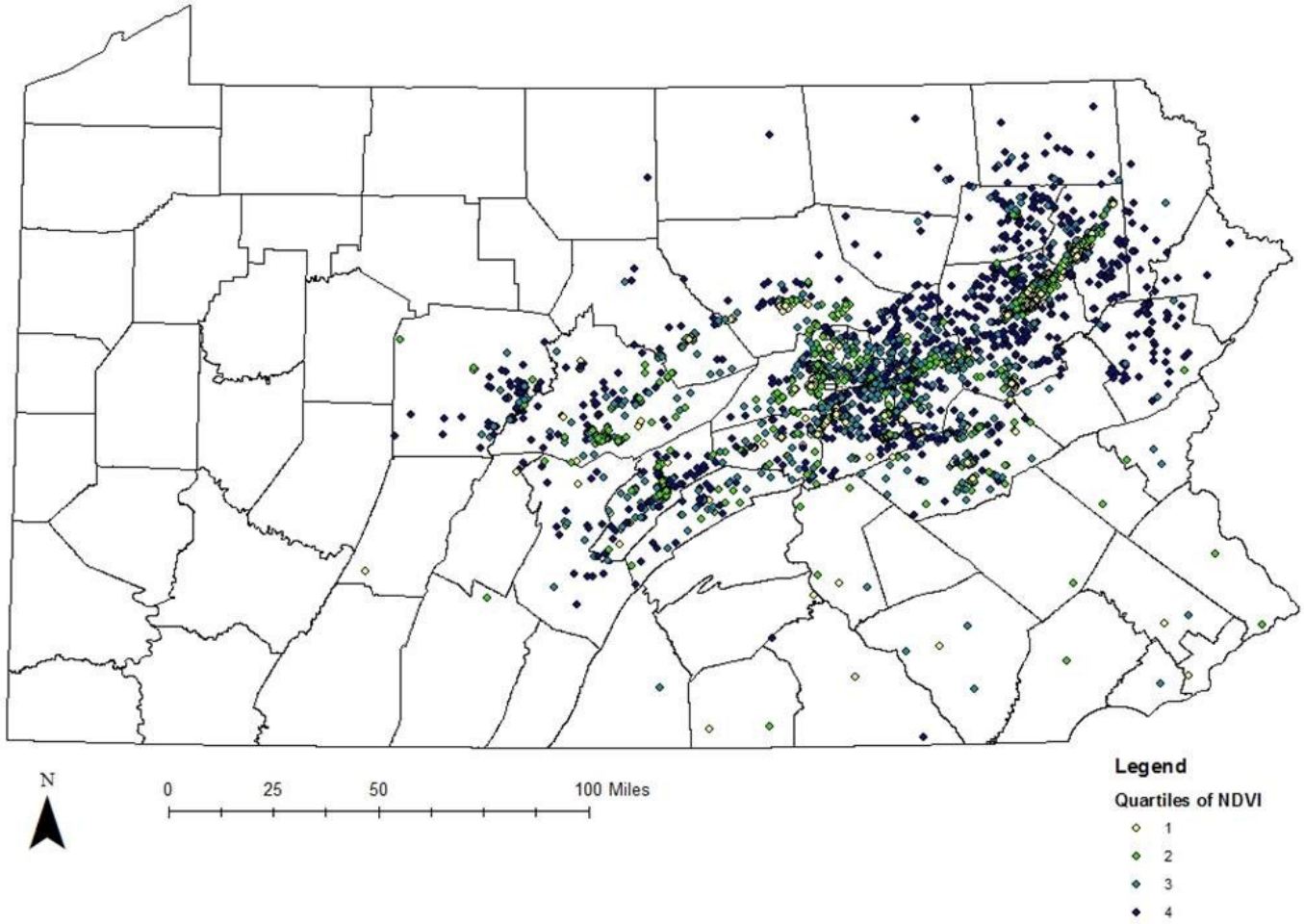


Figure 4.2b.

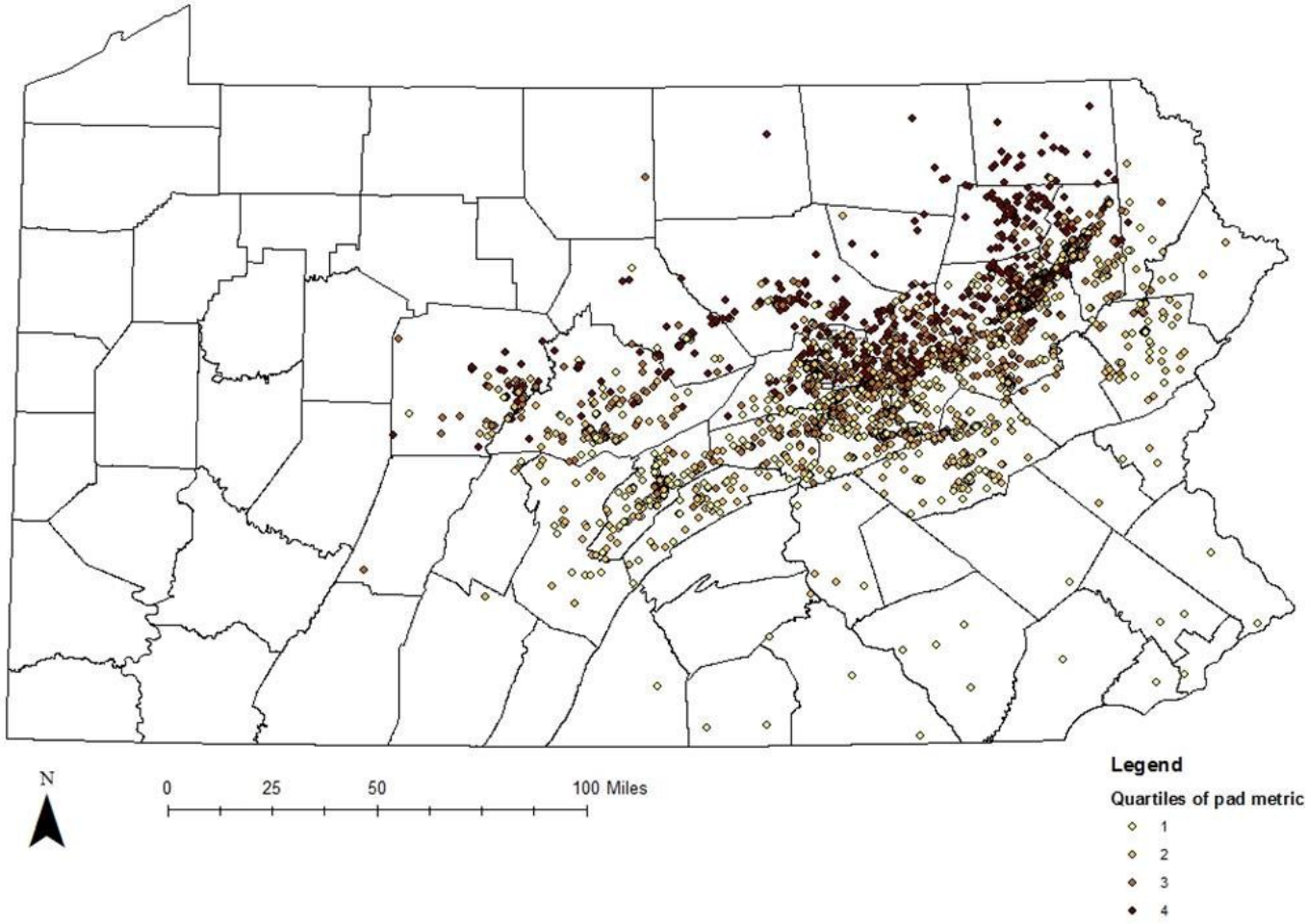


Figure 4.2c.

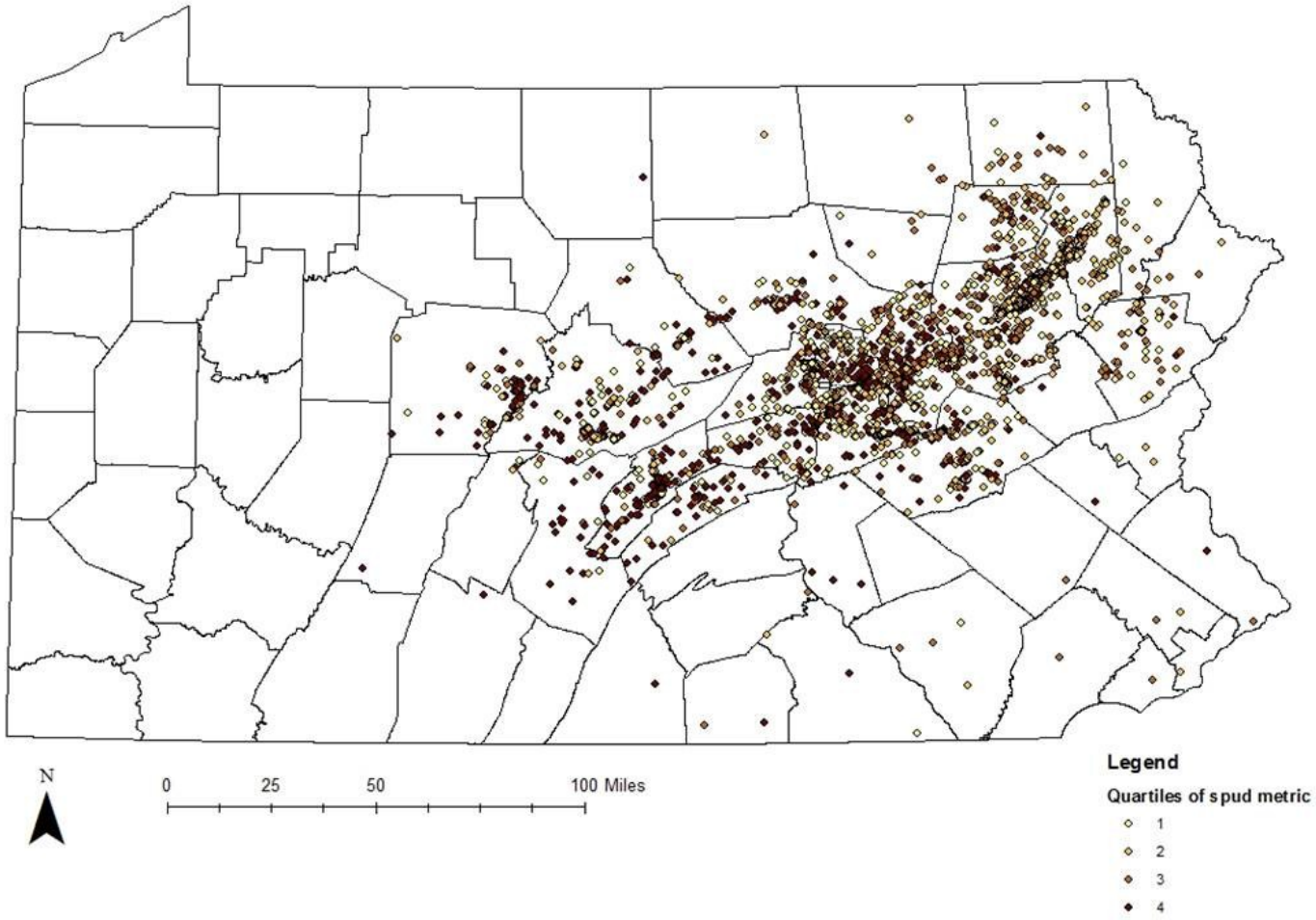


Figure 4.2d.

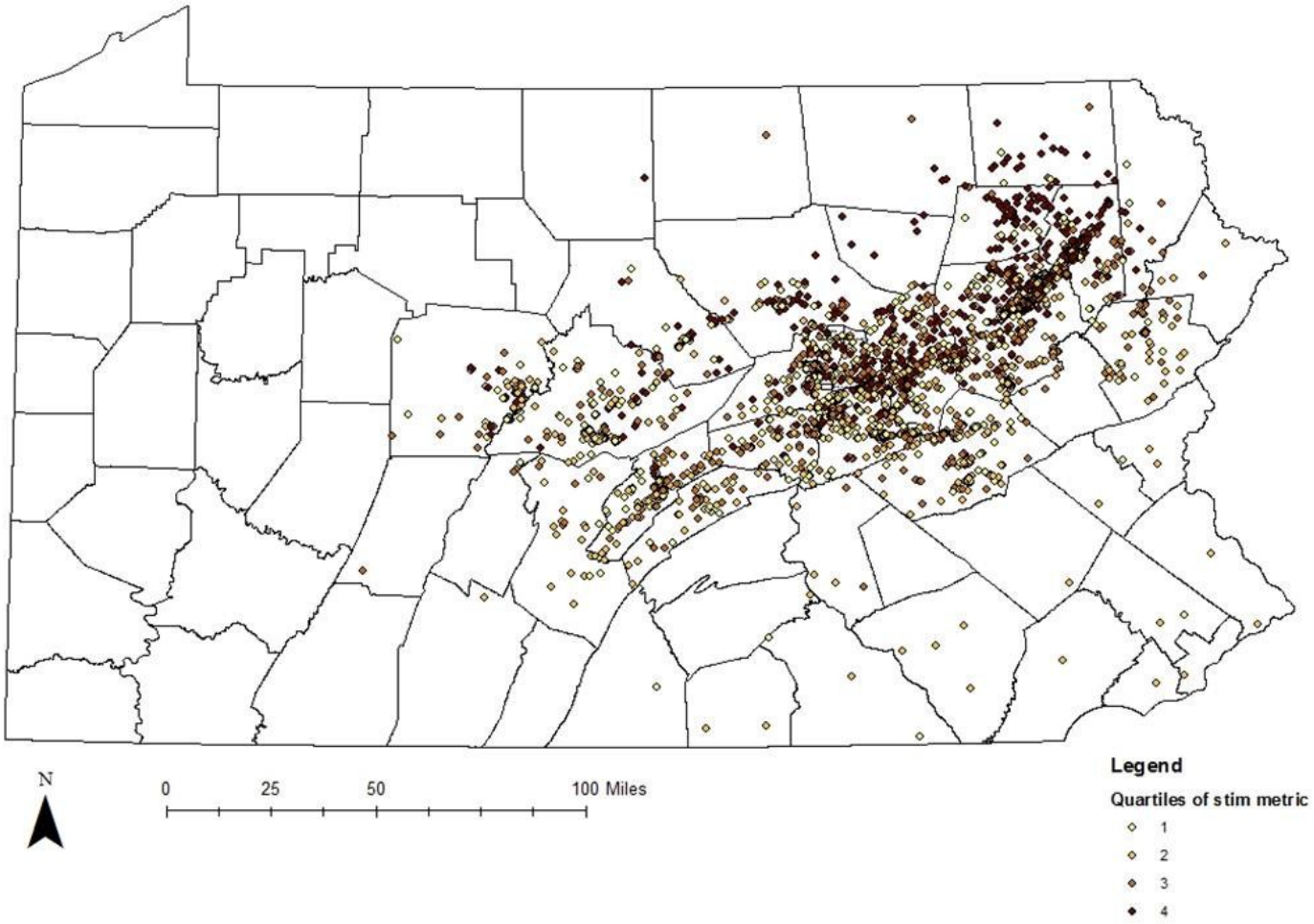


Figure 4.2e.

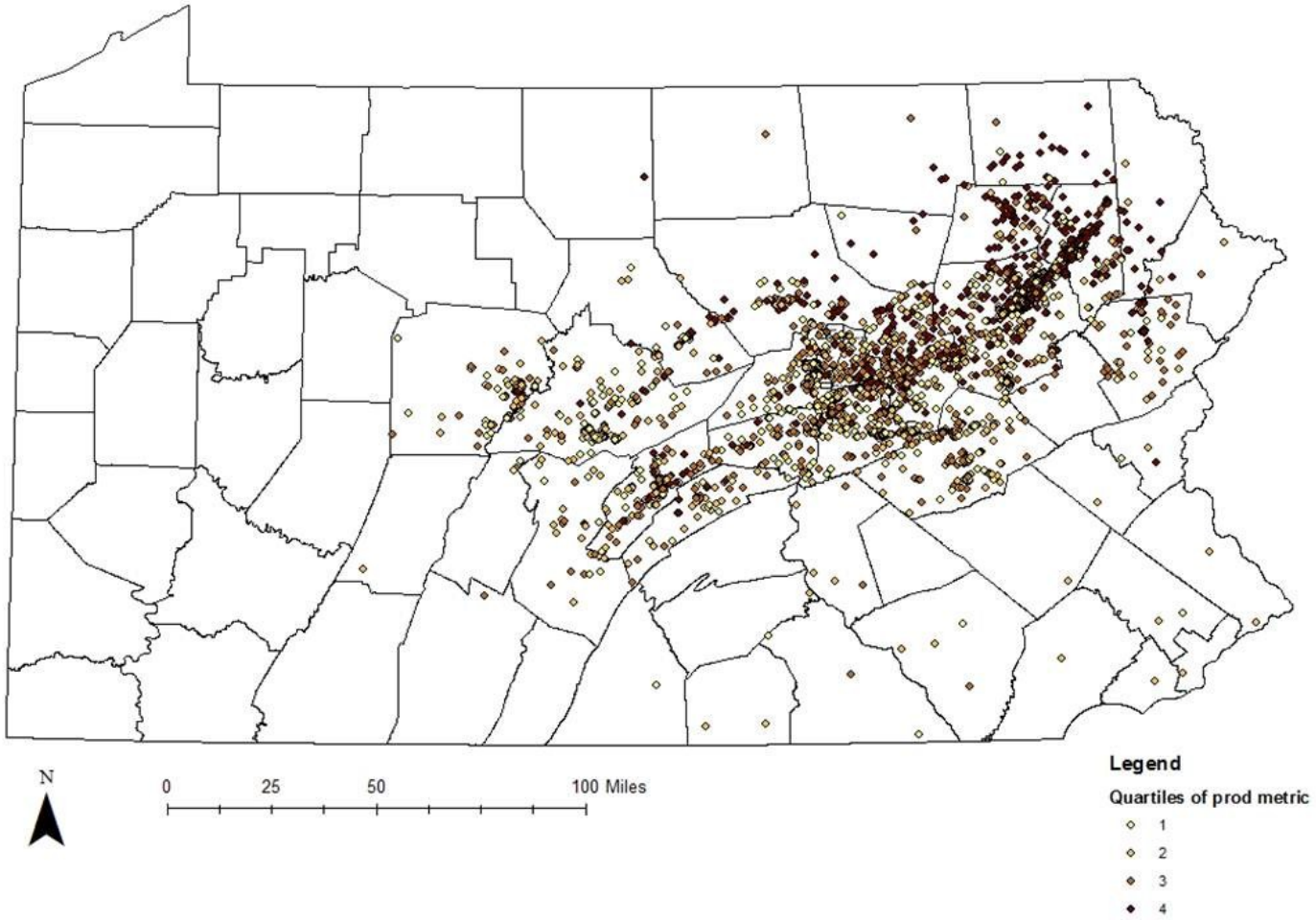


Table 4.5. Adjusted associations (odds ratio [OR], 95% confidence interval [CI]) of unconventional natural gas development (UNGD), normalized difference vegetation index (NDVI), community socioeconomic deprivation (CSD), and distance to roads (in quartiles) with B-type natriuretic peptide (BNP) value ≥ 400 pg/mL (vs. lower)

		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Metric	Quartile	**OR (95 % CI)**	**OR (95 % CI)**	**OR (95 % CI)**	**OR (95 % CI)**	**OR (95 % CI)**	**OR (95 % CI)**
Pad preparation	1	Ref	Ref	Ref	Ref	Ref	Ref
	2	0.97 (0.84, 1.12)	0.98 (0.85, 1.13)	0.98 (0.85, 1.14)	0.97 (0.82, 1.14)	0.98 (0.85, 1.14)	0.97 (0.83, 1.14)
	3	1.09 (0.95, 1.26)	1.11 (0.96, 1.28)	1.12 (0.97, 1.30)	1.09 (0.91, 1.30)	1.15 (0.99, 1.33)	1.12 (0.93, 1.35)
	4	1.04 (0.90, 1.21)	1.06 (0.91, 1.23)	1.08 (0.92, 1.25)	1.02 (0.84, 1.24)	1.10 (0.94, 1.29)	1.06 (0.87, 1.30)
Spud	1	Ref	Ref	Ref	Ref	Ref	Ref
	2	1.15 (1.00, 1.32)	**1.16 (1.01, 1.34)**	**1.17 (1.01, 1.35)**	1.12 (0.91, 1.38)	**1.17 (1.01, 1.35)**	1.11 (0.90, 1.37)
	3	1.04 (0.90, 1.20)	1.05 (0.91, 1.22)	1.06 (0.92, 1.23)	1.04 (0.81, 1.34)	1.07 (0.92, 1.24)	1.02 (0.80, 1.32)
	4	1.10 (0.95, 1.28)	1.12 (0.96, 1.30)	1.15 (0.98, 1.34)	1.15 (0.88, 1.50)	1.13 (0.96, 1.32)	1.10 (0.83, 1.44)
Stimulation	1	Ref	Ref	Ref	Ref	Ref	Ref
	2	1.08 (0.94, 1.24)	1.09 (0.94, 1.25)	1.10 (0.95, 1.27)	1.01 (0.85, 1.20)	1.10 (0.95, 1.26)	1.01 (0.85, 1.21)
	3	1.06 (0.92, 1.23)	1.08 (0.93, 1.25)	1.10 (0.94, 1.27)	0.99 (0.79, 1.24)	1.10 (0.95, 1.28)	1.02 (0.81, 1.27)
	4	1.02 (0.88, 1.18)	1.03 (0.89, 1.19)	1.03 (0.89, 1.20)	0.97 (0.76, 1.23)	1.04 (0.90, 1.22)	1.00 (0.79, 1.28)
Production	1	Ref	Ref	Ref	Ref	Ref	Ref
	2	**1.20 (1.04, 1.37)**	**1.21 (1.06, 1.39)**	**1.24 (1.07, 1.42)**	**1.33 (1.06, 1.68)**	**1.23 (1.07, 1.42)**	**1.36 (1.08, 1.71)**
	3	1.12 (0.97, 1.29)	1.14 (0.98, 1.32)	1.16 (1.00, 1.35)	1.36 (0.99, 1.81)	**1.17 (1.01, 1.36)**	**1.42 (1.05, 1.93)**
	4	1.10 (0.95, 1.28)	1.13 (0.97, 1.31)	1.12 (0.96, 1.31)	1.39 (0.99, 1.96)	1.14 (0.97, 1.34)	**1.52 (1.07, 2.17)**
NDVI*	1	Ref	Ref	Ref	Ref	Ref	Ref
	2	0.97 (0.82, 1.14)	0.97 (0.83, 1.14)	0.97 (0.82, 1.14)	0.98 (0.84, 1.16)	0.96 (0.82, 1.13)	0.97 (0.83, 1.15)
	3	1.01 (0.85, 1.19)	1.01 (0.85, 1.19)	1.01 (0.85, 1.19)	1.04 (0.88, 1.23)	1.00 (0.84, 1.18)	1.03 (0.87, 1.22)
	4	0.90 (0.76, 1.07)	0.90 (0.76, 1.07)	0.89 (0.75, 1.06)	0.93 (0.78, 1.10)	0.88 (0.74, 1.04)	0.91 (0.77, 1.09)
CSD**	1	Ref	Ref	Ref	Ref	Ref	Ref
	2	0.82 (0.78, 1.10)	0.92 (0.77, 1.09)	0.92 (0.77, 1.09)	0.92 (0.77, 1.09)	0.90 (0.75, 1.07)	0.90 (0.75, 1.07)
	3	0.92 (0.78, 1.10)	0.92 (0.77, 1.10)	0.92 (0.77, 1.10)	0.92 (0.77, 1.10)	0.93 (0.78, 1.11)	0.93 (0.78, 1.11)
	4	1.11 (0.93, 1.32)	1.10 (0.92, 1.31)	1.10 (0.92, 1.31)	1.10 (0.92, 1.31)	1.10 (0.92, 1.31)	1.10 (0.93, 1.32)
Distance to major roads (m)†	1	Ref	Ref	Ref	Ref	Ref	Ref
	2	1.06 (0.89, 1.27)	1.06 (0.89, 1.26)	1.07 (0.90, 1.28)	1.07 (0.89, 1.27)	1.08 (0.90, 1.28)	1.07 (0.89, 1.29)
	3	1.09 (0.91, 1.30)	1.09 (0.92, 1.30)	1.10 (0.92, 1.31)	1.10 (0.92, 1.31)	1.09 (0.91, 1.31)	1.09 (0.91, 1.31)
	4	1.00 (0.83, 1.19)	0.99 (0.83, 1.18)	1.00 (0.84, 1.19)	1.00 (0.84, 1.20)	1.00 (0.84, 1.20)	1.00 (0.84, 1.20)

Distance to minor roads (m) [†]	1	Ref	Ref	Ref	Ref	Ref	Ref
	2	0.85 (0.71, 1.01)	**0.84 (0.70, 1.00)**	**0.83 (0.70, 0.99)**	**0.83 (0.70, 0.99)**	**0.83 (0.69, 0.99)**	**0.83 (0.70, 0.99**)
	3	0.93 (0.78, 1.10)	0.92 (0.78, 1.10)	0.92 (0.77, 1.09)	0.92 (0.77, 1.10)	0.91 (0.76, 1.08)	0.91 (0.77, 1.09)
	4	0.91 (0.77, 1.09)	0.91 (0.76, 1.08)	0.91 (0.76, 1.08)	0.91 (0.76, 1.08)	0.88 (0.74, 1.05)	0.88 (0.74, 1.05)

*Normalized difference vegetation index (NDVI), a measure of greenness, was measured by a 250 m x 250 m grid surrounding a subject's residential address for the period of maximum annual greenness in the year of each laboratory measurement

**Community socioeconomic deprivation (CSD), measured from the 2006-2010 US Census American Community Survey at each subject's designated community type.

[†]Distance from each subject's residential address to major and minor roads was calculated in meters using road data from the Federal Highway Administration and ArcGIS 10.4's Generate Near Table function

Model 1: GEE model with an exchangeable correlation matrix, adjusted for sex (female vs. male), nonwhite vs.white, smoking status (ever vs.never), age at time of laboratory date (centered), laboratory setting (inpatient/outpatient), receipt of Medical Assistance (ever vs. never prior to laboratory date), Charlson index (centered and centered squared), body mass index BMI (centered and centered squared), and diagnosis of chronic kidney disease

Model 2: Model 1 + antihyperlipidemic medication and anticoagulant medication

Model 3: Model 2 + duration of HF (date of lab order – the date of first HF diagnosis, included as a centered and centered squared term)

Model 4: Model 3 + year

Model 5: Model 3 + region

Model 6: Model 5 + year

Table 4.6. Summary of the adjusted odds ratios for a B-type natriuretic peptide (BNP) value ≥ 400 pg/mL, by quartile of various environmental metrics, sensitivity analyses with only one observation per person

		Model 1
Metric	Quartile	**OR (95% CI)**
Pad preparation	1	Ref
	2	0.88 (0.71, 1.08)
	3	1.01 (0.82, 1.23)
	4	0.92 (0.75, 1.14)
Spud	1	Ref
	2	1.07 (0.88, 1.31)
	3	0.90 (0.73, 1.10)
	4	0.99 (0.81, 1.22)
Stimulation	1	Ref
	2	1.10 (0.90, 1.34)
	3	0.95 (0.77, 1.16)
	4	0.93 (0.75, 1.14)
Production	1	Ref
	2	1.10 (0.35, 1.33)
	3	1.01 (0.82, 1.23)
	4	1.00 (0.81, 1.24)
NDVI*	1	Ref
	2	0.89 (0.73, 1.09)
	3	0.98 (0.80, 1.20)
	4	**0.80 (0.66, 0.98)**
CSD**	1	Ref
	2	0.88 (0.72, 1.07)
	3	0.93 (0.76, 1.14)
	4	1.10 (0.91, 1.35)
Distance to major roads (m)†	1	Ref
	2	1.02 (0.84, 1.25)
	3	0.97 (0.79, 1.18)
	4	0.93 (0.76, 1.13)
Distance to minor roads (m) †	1	Ref
	2	**0.82 (0.67, 0.99)**
	3	0.88 (0.73, 1.07)
	4	**0.81 (0.66, 0.99)**

*Normalized difference vegetation index (NDVI), a measure of greenness, was measured by a 250 m x 250 m grid surrounding a subject's residential address for the period of maximum annual greenness in the year of each laboratory measurement

**Community socioeconomic deprivation (CSD), measured from the 2010 US Census American Community Survey at each subject's designated place.

†Distance from each subject's residential address to major and minor roads was calculated in meters using road data from the Federal Highway Administration and ArcGIS 10.4's Generate Near Table function

Table 4.7. Adjusted odds ratios for a B-type natriuretic peptide (BNP) value ≥ 400 pg/mL, by quartile of environmental metrics, accounting for native (Model 1) or truncated (Model 2) inverse probability weights for all those potentially eligible for inclusion (n = 6501 laboratory measures among 3938 subjects, weighted to sample of n = 13,183)

		Model 1	**Model 2**
Metric	Quartile	**OR (95% CI)**	**OR (95% CI)**
Pad preparation	1	Ref	Ref
	2	0.97 (0.81, 1.17)	0.97 (0.80, 1.17)
	3	1.06 (0.85, 1.31)	1.05 (0.85, 1.30)
	4	1.01 (0.80, 1.26)	0.99 (0.79, 1.25)
Spud	1	Ref	Ref
	2	1.08 (0.86, 1.37)	1.08 (0.86, 1.37)
	3	0.98 (0.74, 1.30)	0.07 (0.73, 1.28)
	4	1.09 (0.80, 1.48)	1.08 (0.80, 1.48)
Stimulation	1	Ref	Ref
	2	1.07 (0.88, 1.31)	1.07 (0.88, 1.30)
	3	1.03 (0.80, 1.33)	1.04 (0.81, 1.34)
	4	0.99 (0.75, 1.31)	1.00 (0.76, 1.32)
Production	1	Ref	Ref
	2	**1.36 (1.05, 1.76)**	**1.35 (1.04, 1.75)**
	3	**1.48 (1.05, 2.08)**	**1.44 (1.02, 2.02)**
	4	**1.56 (1.05, 2.32)**	**1.53 (1.04, 2.28)**
NDVI*	1	Ref	Ref
	2	0.94 (0.77, 1.14)	0.94 (0.78, 1.13)
	3	0.98 (0.81, 1.19)	0.99 (0.81, 1.20)
	4	0.85 (0.69, 1.04)	0.85 (0.70, 1.04)
CSD**	1	Ref	Ref
	2	0.92 (0.76, 1.13)	0.90 (0.74, 1.09)
	3	0.92 (0.75, 1.12)	0.91 (0.75, 1.11)
	4	1.14 (0.93, 1.39)	1.14 (0.94, 1.38)
Distance to major roads (m)†	1	Ref	Ref
	2	1.03 (0.84, 1.26)	1.02 (0.83, 1.24)
	3	1.06 (0.87, 1.30)	1.04 (0.85, 1.27)
	4	0.95 (0.77, 1.16)	0.93 (0.77, 1.14)
Distance to minor roads (m) †	1	Ref	Ref
	2	0.83 (0.68, 1.02)	0.85 (0.70, 1.03)
	3	0.89 (0.73, 1.09)	0.90 (0.74, 1.10)
	4	0.88 (0.71, 1.08)	0.89 (0.72, 1.08)

*Normalized difference vegetation index (NDVI), a measure of greenness, was measured by a 250 m x 250 m grid surrounding a subject's residential address for the period of maximum annual greenness in the year of each laboratory measurement

**Community socioeconomic deprivation (CSD), measured from the 2010 US Census American Community Survey at each subject's designated place.

†Distance from each subject's residential address to major and minor roads was calculated in meters using road data from the Federal Highway Administration and ArcGIS 10.4's Generate Near Table function

Model 1: GEE model with an exchangeable correlation matrix, adjusts for sex (female vs. male), nonwhite vs. white, smoking status (ever vs. never), age at time of lab (centered), laboratory setting (inpatient vs. outpatient), receipt of Medical Assistance (ever vs. never prior to laboratory date), Charlson index (centered and centered squared), body mass index (BMI) (centered and centered squared), diagnosis of chronic kidney disease, duration of heart failure (date of laboratory order – the date of first heart failure diagnosis, included as a centered and centered squared term), year, region, and covariates used in generating inverse probability weights: myocardial infarction, coronary artery disease, and valve disease

Model 2: Model 1, with inverse probability weights truncated at the 99th percentile

Figure 4.3a.

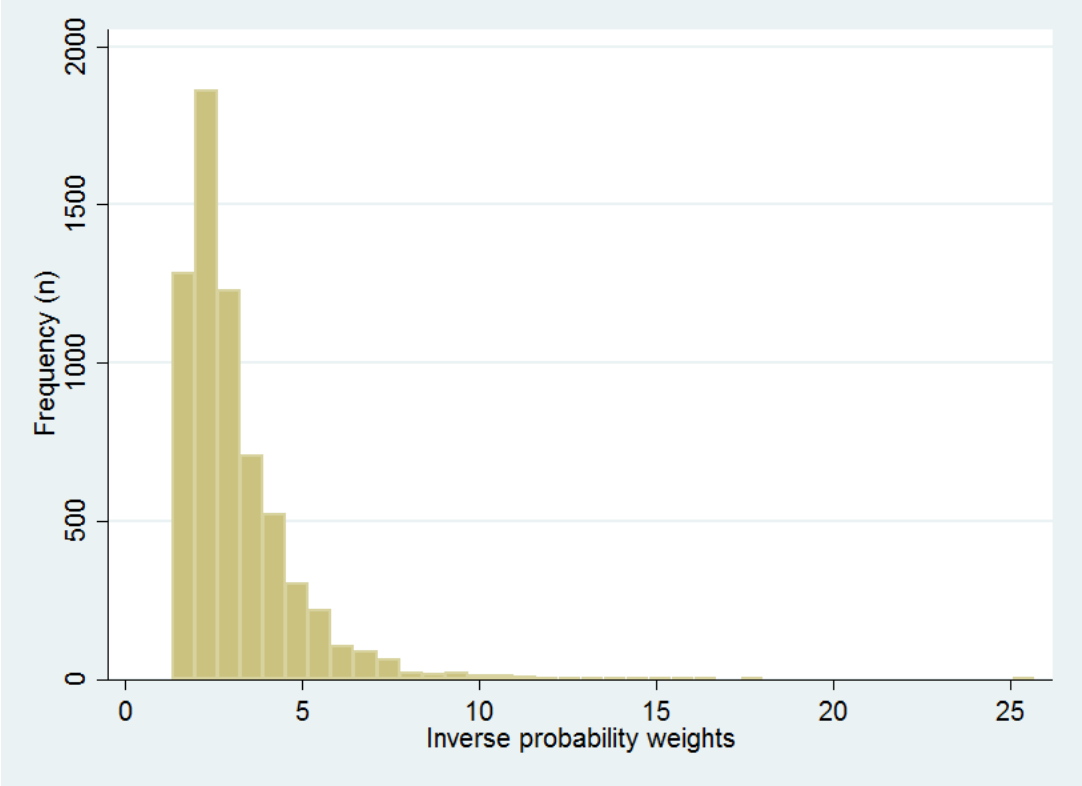

Frequency (n)
0
500
1000
1500
2000
0
5
10
15
20
25
Inverse probability weights

Figure 4.3b.

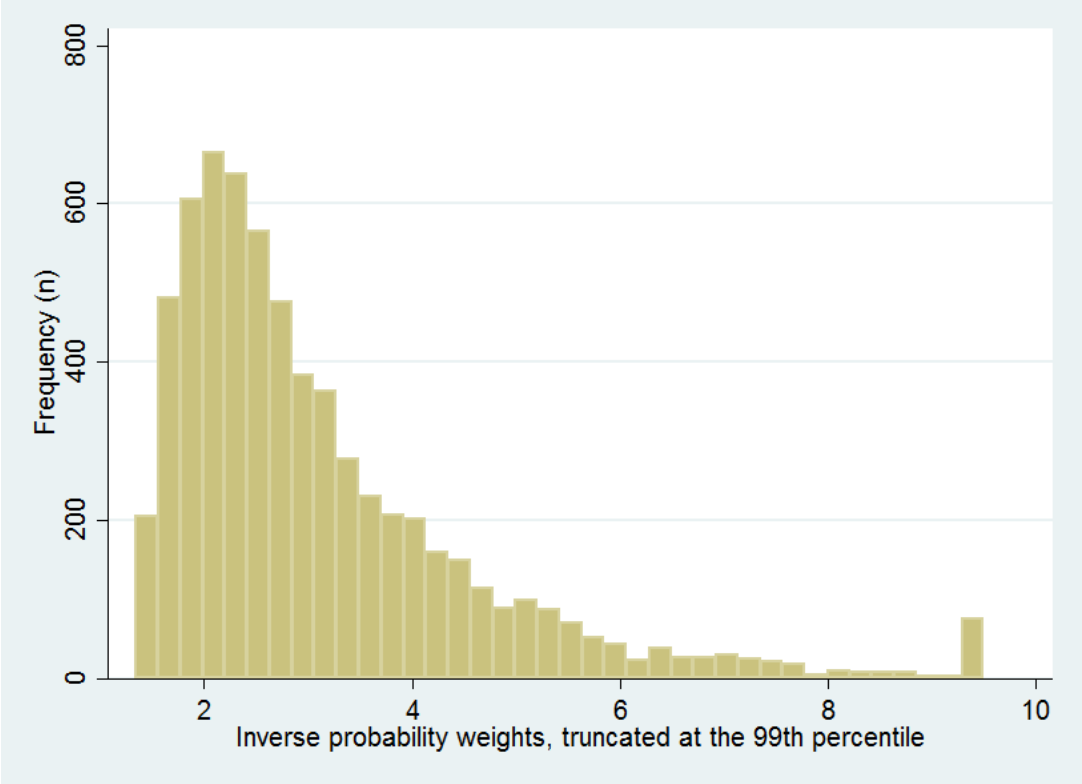

Frequency (n)
0
200
400
600
800
2
4
6
8
10
Inverse probability weights, truncated at the 99th percentile

Table 4.8. Associations (e^{β}*, 95% confidence interval [CI]) of environmental and community metrics, by quartile, from linear regression models using generalized estimating equations of *ln*-B-type natriuretic peptide (BNP) value (pg/mL)

Metric	Quartile	$e^{\beta*}$ (95% CI)
Pad preparation	1	Ref
	2	1.02 (0.94, 1.11)
	3	**1.11 (1.01, 1.21)**
	4	1.05 (0.95, 1.15)
Spud	1	Ref
	2	**1.15 (1.04, 1.28)**
	3	1.11 (0.98, 1.26)
	4	1.13 (0.99, 1.29)
Stimulation	1	Ref
	2	1.04 (0.95, 1.13)
	3	1.02 (0.92, 1.14)
	4	1.07 (0.94, 1.20)
Production	1	Ref
	2	**1.28 (1.15, 1.43)**
	3	**1.30 (1.12, 1.51)**
	4	**1.39 (1.17, 1.65)**
NDVI**	1	Ref
	2	1.00 (0.92, 1.09)
	3	1.01 (0.92, 1.10)
	4	0.94 (0.85, 1.03)
CSD***	1	Ref
	2	1.00 (0.90, 1.11)
	3	1.00 (0.90, 1.10)
	4	1.08 (0.98, 1.20)
Distance to major roads (m)†	1	Ref
	2	1.07 (0.96, 1.18)
	3	1.06 (0.96, 1.18)
	4	1.00 (0.90, 1.11)
Distance to minor roads (m) †	1	Ref
	2	**0.90 (0.81, 0.99)**
	3	0.92 (0.83, 1.02)
	4	0.91 (0.82, 1.00)

* e^{β} represents the ratio of the geometric means of B-type natriuretic peptide (BNP), comparing upper quartiles to the reference quartile

**Normalized difference vegetation index (NDVI), a measure of greenness, was measured by a 250 m x 250 m grid surrounding a subject's residential address for the period of maximum annual greenness in the year of each laboratory measurement

***Community socioeconomic deprivation (CSD), measured from the 2010 US Census American Community Survey at each subject's designated place.

†Distance from each subject's residential address to major and minor roads was calculated in meters using road data from the Federal Highway Administration and ArcGIS 10.4's Generate Near Table function

Chapter 5a: Examination of the association between UNGD activity and heart failure hospitalization by heart failure phenotype status

5A.1 ABSTRACT

Background: In **Chapter 3**, we observed that higher quartiles of three UNGD activity metrics (pad preparation, stimulation, and production) were associated with greater odds of hospitalization and there was evidence of exposure-effect relations with increasing quartile of UNGD activity. In this chapter, given that heart failure pathophysiology is known to differ by phenotypic status (i.e., heart failure with reduced or preserved ejection fraction [HF*r*EF or HF*p*EF]), we evaluated effect modification by heart failure phenotype on the associations of UNGD with heart failure hospitalization.

Methods: We used eMERGE heart failure phenotypes that had been previously assigned to the subjects in **Chapter 3** and evaluated differences in demographics, comorbidities, and medication use by phenotype categories: not phenotyped (including "eMERGE not applied," and eMERGE applied but not able to phenotype ["eMERGE no phenotype"] groups), HF*r*EF, and HF*p*EF subjects. We evaluated effect modification by these phenotype categories on associations of each UNGD activity metric with heart failure hospitalization, by adding cross-product terms to the final adjusted models used in **Chapter 3**. We first hypothesized that that associations of UNGD with heart failure hospitalization would be stronger in subjects with HF*p*EF (vs. HF*r*EF). We then completed a second *post hoc* analysis based on observations from prior chapters and analysis herein that phenotyped subjects (vs. not phenotyped) had a higher prevalence of comorbidities and medication use; shorter duration of disease; and greater risk of hospitalization and death. We thus hypothesized that having any phenotype (i.e., HF*r*EF

and HF*p*EF vs. subjects in the two groups without phenotypes) could be used as a surrogate measure of heart failure severity, and we evaluated effect modification by heart failure severity on the associations of UNGD with hospitalization.

Results: We observed increased odds of hospitalization among the HF*r*EF subjects (OR [95% CI] = 1.16 [1.00, 1.35]) and decreased odds of hospitalization among the eMERGE not applied (0.39 [0.34, 0.44] and no phenotype groups (0.62 [0.53, 0.74]), all compared to HF*p*EF. We observed, in models of the pad metric only, the global test of significance of the cross-products between HF*r*EF and UNGD metric quartiles was p = 0.03. For the pad, spud, and production metrics, the global test of significance for the cross-products between each metric and the eMERGE not applied group was p = 0.0002, p = 0.03, and p = 0.04, respectively. In *post hoc* analyses, an indicator of heart failure severity modified associations between UNGD activity and odds of hospitalization for the pad preparation and spud metrics. Although we did not observe this association in the stim or production metrics, stratum-specific odds ratios reflected trends of increasing odds of hospitalization with increasing quartile of UNGD activity in both heart failure severity groups.

Conclusions: HF*r*EF subjects, compared to HF*p*EF subjects and subjects without a phenotype, had higher odds of hospitalization independent of UNGD activity. Associations of UNGD activity with heart failure hospitalization did not differ by the two primary heart failure phenotypes. However, results from *post hoc* analyses indicated that disease severity modified associations between UNGD activity and hospitalization, consistent with the hypothesis that the study may have identified an important group of older subjects particularly vulnerable to adverse health impacts from UNGD activity.

5A.2 INTRODUCTION

UNGD activity has been associated with a number of health outcomes, which now includes greater odds of hospitalization among heart failure patients (**Chapter 3**). Within the context of heart failure epidemiology, studies that evaluated differences between HF*r*EF and HF*p*EF phenotypes are scarce, but sorely needed [209, 259]. Specifically, HF*p*EF subjects are more likely to die from related comorbidities than from cardiovascular disease itself [261]. Yet, few studies have examined environmental factors or psychosocial stressors in relation to heart failure hospitalization by phenotype [87, 89] despite evidence for phenotypic differences in related risk factors [81, 82, 85, 262]. Understanding associations between UNGD activity and heart failure hospitalizations by phenotype would therefore address a need in the environmental epidemiology of heart failure and add to the epidemiologic literature on UNGD and health outcomes, which to date has primarily studied health outcomes in early to mid-life and notably has not studied UNGD in relation to health conditions that are more prevalent in older persons [9, 11, 13, 16, 64].

Examining the relations between UNGD and heart failure hospitalization by phenotype could also provide insight into potential exposure pathways involved in UNGD-health associations. We hypothesized that associations between UNGD activity metrics and heart failure hospitalizations would be stronger among those with HF*p*EF (vs. HF*r*EF) because the community and environmental impacts of UNGD are related to risk factors and exposures (e.g., obesity, diabetes, high blood pressure and hypertension, air pollution with oxides of nitrogen [NO_2]) that are more strongly associated with HF*p*EF than HF*r*EF [81, 262, 263].

5A.3 METHODS

We used eMERGE heart failure phenotypes that had been previously assigned [179, 185] to the subjects in our hospitalization study (**Chapter 3)**, which was able to

categorize case and control subjects into four phenotypic groups: HF*p*EF, HF*r*EF, those without enough information to be phenotyped ("eMERGE not applied"), and those who had information to have the eMERGE algorithm applied but did not have a clear HF*p*EF or HF*r*EF phenotype ("eMERGE no phenotype"). Greater detail on the eMERGE heart failure phenotyping algorithm can be found in **Chapter 2, section 2.2** and in **Appendix A. Heart failure algorithm**). In analyses in **Chapter 2** and herein, we observed that subjects in the hospitalization study who had the eMERGE algorithm applied to their EHR data had a greater proportion of subjects who were deceased by the end of the study period compared to those who did not have the eMERGE algorithm applied (**Table 2.10**). We also observed that subjects in the hospitalization study who had the eMERGE algorithm applied had a higher proportion of antihypertensive, antihyperlipidemic, and anticoagulant medications and a higher proportion of comorbidities (i.e., chronic obstructive pulmonary disease, coronary artery disease, hypertension, myocardial infarction, valve disorder, type 2 diabetes, chronic kidney disease) and a greater mean Charlson index of morbidity value compared to subjects who did not have the eMERGE algorithm applied (**Table 2.14**). Finally, in model building for the analysis presented in this chapter, we observed that persons with HF*r*EF were more likely to be hospitalized for heart failure. Taken together, we hypothesized that phenotyped subjects had more severe heart failure and used phenotype status as a surrogate measure of the severity of the disease.

We first evaluated differences in patient characteristics (i.e., demographics, comorbidities, medications) between subjects who were not able to be phenotyped, those who were phenotyped as HF*r*EF, and those who were phenotyped as HF*p*EF. We determined that the ability to be phenotyped was dependent on observation time, disease severity, and the presence of comorbidities, so we did not attempt to impute

phenotype information for subjects who were not able to be phenotyped by the eMERGE algorithm.

In our primary analysis, we used the phenotypic categories (eMERGE not applied, eMERGE no phenotype, HF*p*EF, and HF*r*EF) to create three binary indicators, designating HF*p*EF as the reference group. We first assessed the association between each phenotype indicator and hospitalization, independent of UNGD activity in adjusted multilevel logistic regression models (from **Chapter 3**). We next examined adjusted multilevel logistic regression models of hospitalization (from **Chapter 3**) that included the main effects of each phenotype indicator and quartiles of UNGD activity. Lastly, we evaluated models that included cross-product terms between these three indicators and each UNGD metric quartile indicator. To assess whether the inclusion of these cross-products improved the model, we calculated the global p-value of these cross-product terms with chi^2 tests, which tested the hypothesis that the coefficients for each cross-product were equal to zero. We generated linear combinations of both main effects (i.e., main effect of UNGD, main effect of phenotype groups) and the cross-products between these to estimate stratum-specific odds ratios for hospitalization. Using the "ggplot2" package in R v.3.4.2 [264], we generated forest plots to visually display these odds ratios by phenotype group and by UNGD quartile.

In *post hoc* analyses, we assessed effect modification by heart failure severity (i.e., an indicator for having a phenotype [either HF*p*EF or HF*r*EF] vs. no phenotype) on the associations between UNGD metrics and hospitalization. We used the same multilevel logistic regression models in our primary analysis (and from **Chapter 3**) and also included cross-products between each UNGD activity metric and the heart failure severity indicator. We calculated the global p-values of these cross-product terms with chi^2 tests, which tested the null hypothesis that the coefficients for any one of the cross-products were equal to zero. To understand the association between UNGD activity

metrics and odds of hospitalization, after accounting for the independent effects of the heart failure severity indicator, we generated linear combinations of both main effects (i.e., main effect of UNGD, main effect of severity indicator) and the cross-products between these to estimate stratum-specific odds ratios for hospitalization, which we also displayed in forest plots generated in R v.3.4.2 [264]. We conducted chi^2 tests on the inclusion of the cross-product terms between the binary indicator for having a phenotype and each UNGD metric.

5A.4 Results

5A.4.1 Descriptive characteristics of phenotype groups

Comparing subjects by phenotype groups, a greater proportion of subjects with a HF*p*EF phenotype were female (59.8%), whereas a greater proportion of HF*r*EF subjects were male (63.7%); in the group of subjects without a phenotype, the distribution of sex was 52.2% male and 47.8% female (**Table 5a.1**). Subjects with the HF*p*EF phenotype were, on average, older (mean age = 74 years compared to 70 years for HF*r*EF and 71 years for the subjects without a phenotype, **Table 5a.1**). The HF*p*EF group also had the highest proportion of subjects with hypertension, diabetes, chronic obstructive pulmonary disease, and chronic kidney disease (**Table 5a.2**). Meanwhile, subjects with HF*r*EF had a higher prevalence of coronary artery disease and of having had a myocardial infarction compared to HF*p*EF subjects and those without a phenotype (**Table 5a.2**). Comparing these two phenotype groups to the subjects without a phenotype, we observed that subjects with a phenotype had a greater proportion of medication use (antihypertensive, antihyperlipidemic, and anticoagulant), comorbidities (chronic obstructive pulmonary disease, coronary artery disease, hypertension, valve disorder, diabetes, and chronic kidney disease), and subjects who were deceased by the end of the study period (39% of subjects were deceased in the phenotyped groups, compared to 30% among the no phenotype group) (**Tables 5a.1-2**). We also noticed that

subjects who were not phenotyped had a longer mean duration of heart failure compared to those who had a phenotype (855 days vs. 754 and 670 days, $p < 0.0001$), suggesting less severe disease among those without phenotypes, because severe heart failure is associated with worse outcomes and thus shorter duration of disease (**Table 5a.1**).

5A.4.2 Associations of phenotype with hospitalization and with UNGD activity

We first evaluated phenotype in relation to heart failure hospitalization without UNGD activity in the models and found that HF*r*EF was associated with greater odds of hospitalization (OR [95 % CI] = 1.16 [1.00, 1.35]) compared to HF*p*EF; the other phenotype groups had a reduced odds of hospitalization (OR [95 % CI] = 0.62 [0.53, 0.74]) for the no phenotype group and 0.39 [0.34, 0.44] for the eMERGE not applied group) compared to HF*p*EF (**Table 5a.3**). UNGD activity was next added to these models and we found some evidence of confounding by phenotype, as associations of UNGD activity with hospitalization were slightly attenuated, although previously described associations of pad preparation, stimulation, and production metrics with hospitalization were still present (**Table 5a.4**), with significantly higher odds of hospitalization for quartiles 2 - 4 of the pad metric compared to the first quartile (OR [95% CI] = 1.16 [1.01, 1.33], 1.52 [1.31, 1.78], and 1.57 [1.30, 1.89]). For the stimulation metric, we saw higher odds for the 3rd and 4th quartiles (1.42 [1.14, 1.77] and 1.60 [1.26, 2.02], respectively); and for the production metric, we only saw a 4th quartile association (1.55 [1.11, 2.17]). These phenotype-adjusted associations of UNGD activity with hospitalization exhibited similar exposure-effect and 4th quartile associations as did the associations reported in **Chapter 3**, where associations in quartiles 2 – 4 of the pad metric compared to the first quartile were OR (95% CI) = 1.19 (1.01, 1.40), 1.63 (1.35, 1.97), and 1.70 (1.35, 2.13). Similarly the associations in the 3rd and 4th quartiles of the

stim metric (from **Chapter 3**) were 1.56 (1.19, 2.04), and 1.80 (1.35, 2.40), and the 4th quartile effect was present for the production metric 1.62 (1.07, 2.45).

5A.4.3 Effect modification by phenotype on associations of UNGD activity with hospitalization

In models that evaluated cross-products between these phenotypic indicators and UNGD metrics, we observed that HF*r*EF phenotype modified associations of the pad metric with hospitalization (p = 0.03 for global test of three cross-products, **Table 5a.5**), although this effect modification was primarily driven by cross-products of the 2nd quartiles of the pad preparation metric rather than the cross-products of the 3rd and 4th quartiles of the pad preparation metric. This could suggest that effect modification is only present in lower levels of UNGD activity, or that these associations are spurious. We did not observe effect modification by HF*r*EF (vs. HF*p*EF) for the other UNGD metrics (**Tables 5a.6** – **5a.8**). However, we observed effect modification by eMERGE not applied (part of the not phenotyped, vs. HF*p*EF) for the pad metric (p = 0.0002 for global test of three cross-products, **Table 5a.5**), spud metric (p = 0.03 for global test of three cross-products, **Table 5a.6**), and production metric (p = 0.04 for global test of three cross-products, **Table 5a.8**).

From the models that incorporated cross-products between phenotypic indicators and UNGD metrics (i.e., model results displayed in **Tables 5a.5-8**), we were able to estimate stratum-specific odds ratios of hospitalization by UNGD activity metrics and heart failure phenotype groups (**Figures 5a.1-4**). The figures display the linear combinations of the stratum-specific odds ratio of hospitalization for each UNGD activity metric, as the combined odds of the phenotype group, UNGD activity metric quartiles, and their cross-products. The forest plots show that the associations of UNGD with heart failure hospitalization were elevated but relatively similar for the HF*p*EF and HF*r*EF groups, but reduced and relatively similar for the not phenotyped groups. These results

did not provide evidence that the associations of UNGD with heart failure hospitalization differed directly comparing HF*p*EF and HF*r*EF, our primary *a priori* hypothesis of interest.

5A.4.4 *Post hoc* analysis: effect modification by heart failure severity on relations of UNGD activity with hospitalization

In *post hoc* analyses, there was evidence that heart failure severity (phenotyped vs. no phenotype) modified relations of UNGD activity metrics with hospitalization. The cross-products between this severity indicator and spud metric quartiles were associated with increased odds of hospitalization (OR [95 % CI]) in third quartile = 1.30 [1.01, 1.68]; fourth quartile = 1.34 [1.03, 1.73]), and the global test of significance for these cross-products was p = 0.009 (**Table 5a.9**). The cross-products between quartiles of the pad metric and the severity indicator were associated with reduced odds of hospitalization in the second (vs. first) quartile (OR [95 % CI] = 0.71 [0.55, 0.23]), but not in the third or fourth quartiles, although the global test of significance for this cross product was p = 0.03 (**Table 5a.9**). Although the second (vs. first) quartile of the cross-product of our severity indicator and the production metric was associated with increased odds of hospitalization (OR [95 % CI] = 1.40 [1.08, 1.80]), we observed null associations in the third and fourth quartiles of these cross-products, and the global test of significance for these cross-products was p = 0.08 (**Table 5a.9**). We also observed null associations for the cross-products between quartiles of the stimulation metric and the severity indicator, and the global test of significance for these terms was p = 0.4 (**Table 5a.9**).

To observe the combined effect of each UNGD activity metric, severity indicator, and cross-products between the severity indicator and each UNGD activity metric, we generated forest plots from the combined associations with odds of hospitalization. We display these for each UNGD activity metric in **Figures 5a.5-8**. There was evidence that the severe heart failure indicators modified relations of both the pad (**Figure 5a.5**) and

spud (**Figure 5a.6**) UNGD activity metrics with hospitalization. Although the associations of the stimulation metric (**Figure 5a.7**) and the production metric (**Figure 5a.8**) were stronger among the more severe subjects and increased with increasing quartile of UNGD activity, the global tests of significance of the cross-products suggested that the trends observed were due to the combined main effects of each metric and the severity indicator, not a statistical interaction between the two.

5A.5 DISCUSSION

The purpose of this analysis was to evaluate the association between heart failure phenotype with heart failure hospitalization, and to evaluate if heart failure phenotype modified the previously observed (**Chapter 3**) associations between UNGD activity and heart failure hospitalization. We also observed that the HF*r*EF indicator was associated with increased odds of hospitalization, and that not phenotyped groups were associated with reduced odds of hospitalization. These main effects, in addition to descriptive analyses comparing phenotyped subjects to those who were not phenotyped, led us to use phenotyped status as a surrogate for heart failure severity. This was further supported by the greater prevalence of comorbidities, medication use, and a higher proportion of subjects who became deceased by the end of the study period among those with either a HF*r*EF or HF*p*EF phenotype compared to those without a phenotype. When we incorporated the main effects of phenotype indicators (i.e., HF*r*EF, eMERGE no phenotype, and eMERGE not applied vs. HF*p*EF), the main effects of UNGD activity metrics, and the cross-products for each of these indicators with each respective UNGD activity metric, we did not find evidence that our *a priori* hypothesis was correct (i.e., stronger associations among HF*p*EF vs. HF*r*EF subjects). However, we did find, in *post hoc* analyses, that phenotyped subjects with either HF*p*EF or HF*r*EF, that is, more

severe heart failure, evidenced stronger associations of UNGD with heart failure hospitalization than did those with less severe heart failure.

The results of this analysis support the findings of **Chapter 3**, since we also observed associations of the UNGD pad preparation metric, the stimulation metric, and the production metric with hospitalization for heart failure. However, in our phenotype-adjusted models, we saw a small attenuation of the associations of each UNGD activity metric compared to our findings in **Chapter 3**. This suggested that phenotype could be a confounder of the association between UNGD activity and heart failure hospitalization. When we included cross-products between UNGD activity metrics and our heart failure severity indicator (i.e., HF*p*EF or HF*r*EF) in *post hoc* analyses, we observed that heart failure severity modified relations of the UNGD spud metric with heart failure hospitalization, which was of particular interest because we did not observe associations between the spud metric and hospitalization in **Chapter 3**, except for in a sensitivity analysis when we used 4:1 control to case frequency matching. This association suggests that, although not observed in our original analysis, the spud metric may be associated with heart failure hospitalization among the most severe cases.

Even though we were not able to assign all subjects a phenotype category with the eMERGE algorithm, we consider the successful application of the phenotyping algorithm an informative finding. We observed that subjects with more frequent contact with the health care system; who had more comorbidities and medication use; a shorter duration of disease; and were more likely to be deceased by the end of the study period, were more likely to have an assigned phenotype by the eMERGE algorithm, and they had a greater likelihood of hospitalization compared to those who were not assigned a phenotype with the algorithm. Although the eMERGE algorithm was developed in 2015 [179], we have not been able to identify any environmental epidemiology studies that utilize this method in a large population.

We have not been able to identify any epidemiologic studies that have evaluated associations between any environmental variables and heart failure hospitalization that evaluated effect modification by HF*p*EF and HF*r*EF phenotypes or by disease severity. One study in Spain identified 353 individuals admitted to a tertiary care hospital with a diagnosis of heart failure and found that, comparing individuals with HF*p*EF to those with HF*r*EF, those with HF*p*EF had higher levels of ambient NO_2 in the previous week before their hospital admission [87]. Another study of heart failure hospitalizations in Medicare beneficiaries in Pittsburgh, Pennsylvania found that air pollutants (carbon monoxide [CO], particulate matter less than 2.5 microns in diameter [$PM_{2.5}$], oxides of nitrogen [NO_2] and oxides of sulfur [SO_2]) were associated with increased rates of hospitalization and that having a recent myocardial infarction diagnosis modified the association between $PM_{2.5}$ and hospitalization rate [130]. However, these two studies did not specifically evaluate whether heart failure phenotype modified this association, nor did they evaluate indicators of heart failure severity as effect measure modifiers of the air pollution and hospitalization associations.

Strengths of this study include the large sample size from EHR data, from subjects who are representative of the general population in Pennsylvania [182]. This data source, coupled with the novel advantages of the eMERGE heart failure phenotype algorithm [179, 185], allowed us to distinguish subjects into phenotype groups with a high positive predictive value (> 95%) [179]. The validity of the phenotype groups is supported by our observations of characteristics and comorbidities associated with HF*p*EF and HF*r*EF phenotypes that are consistent with the epidemiologic literature on HF*p*EF and HF*r*EF; we found that a greater proportion of HF*p*EF subjects were female and had comorbidities such as type 2 diabetes, chronic obstructive pulmonary disease, hypertension, and chronic kidney disease, whereas HF*r*EF subjects had a higher proportion of coronary artery disease and a previous myocardial infarction diagnosis

[263, 265]. Further, this study is the first to have applied the eMERGE heart failure algorithm in an environmental epidemiology study of heart failure subjects with different phenotypes.

Limitations of this study include the inability to measure dietary, behavioral, and occupational factors through the EHR, which is a limitation that is common to EHR-based epidemiology studies. Another limitation is that not all subjects were able to be phenotyped with the eMERGE algorithm because the algorithm required echocardiogram measures and consistent contact with the health care system (**Appendix A. Heart failure algorithm**). However, we learned that the presence of a phenotype was informative as an indicator of disease severity, which might have applications to other epidemiology studies that utilize EHR-based phenotyping algorithms.

5A.6 Conclusions

We observed that phenotyped subjects, with HF*p*EF or HF*r*EF, had a greater odds of hospitalization compared to the no phenotype and eMERGE not applied groups. We initially hypothesized that the HF*p*EF phenotype (vs. HF*r*EF) would modify the association between UNGD activity and hospitalization but did not find evidence of this. However, we did observe that disease severity, based on the results of the application of the eMERGE algorithm to EHR data (phenotyped vs. not phenotyped), modified the association between UNGD activity and hospitalization that we observed in **Chapter 3**, particularly for the pad preparation and spud metrics.

5A.7 REFERENCES

1. von Lueder, T.G. and S. Agewall, *The burden of heart failure in the general population: a clearer and more concerning picture.* Journal of thoracic disease, 2018. **10**(Suppl 17): p. S1934-S1937.
2. Conrad, N., et al., *Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals.* The Lancet, 2018. **391**(10120): p. 572-580.
3. Gerber, Y., et al., *A contemporary appraisal of the heart failure epidemic in Olmsted County, Minnesota, 2000 to 2010.* (2168-6114 (Electronic)).
4. Dominguez-Rodriguez, A., et al., *Air pollution and heart failure: Relationship with the ejection fraction.* World Journal of Cardiology, 2013. **5**(3): p. 49-53.
5. Bui, A.L., T.B. Horwich, and G.C. Fonarow, *Epidemiology and risk profile of heart failure.* Nature reviews. Cardiology, 2011. **8**(1): p. 30-41.
6. Borlaug, B.A., *Heart failure with preserved and reduced ejection fraction: different risk profiles for different diseases.* European Heart Journal, 2013.
7. Lee, D.S., et al., *Relation of Disease Pathogenesis and Risk Factors to Heart Failure With Preserved or Reduced Ejection Fraction.* Circulation, 2009. **119**(24): p. 3070-3077.
8. Brouwers, F.P., et al., *Incidence and epidemiology of new onset heart failure with preserved vs. reduced ejection fraction in a community-based cohort: 11-year follow-up of PREVEND.* European Heart Journal, 2013. **34**(19): p. 1424-1431.
9. Eaton, C.B., et al., *Risk Factors for Incident Hospitalized Heart Failure With Preserved Versus Reduced Ejection Fraction in a Multiracial Cohort of Postmenopausal Women. LID - e002883 [pii].* (1941-3297 (Electronic)).
10. McKenzie, L.M., et al., *Birth outcomes and maternal residential proximity to natural gas development in rural Colorado.* Environmental health perspectives, 2014. **122**(4): p. 412-417.
11. Stacy, S., et al., *Perinatal outcomes and unconventional natural gas operations in Southwest Pennsylvania.* PLoS One, 2015. **10**(6): p. e0126425.
12. Willis, M.D., et al., *Unconventional natural gas development and pediatric asthma hospitalizations in Pennsylvania.* (1096-0953 (Electronic)).
13. Whitworth, K.W., A.K. Marshall, and E. Symanski, *Maternal residential proximity to unconventional gas development and perinatal outcomes among a diverse urban population in Texas.* PloS one, 2017. **12**(7): p. e0180966-e0180966.
14. Casey, J.A., et al., *Unconventional Natural Gas Development and Birth Outcomes in Pennsylvania, USA*. Epidemiology, 2015.
15. Tsao, C.W., et al., *Temporal Trends in the Incidence of and Mortality Associated With Heart Failure With Preserved and Reduced Ejection Fraction.* (2213-1787 (Electronic)).
16. Bielinski, S.J., et al., *A Robust e-Epidemiology Tool in Phenotyping Heart Failure with Differentiation for Preserved and Reduced Ejection Fraction: the Electronic Medical Records and Genomics (eMERGE) Network.* 2015(1937-5395 (Electronic)).
17. Bielinski;, S.J., *Heart Failure (HF) with Differentiation between Preserved and Reduced Ejection Fraction*, in *PheKB*, M. Clinic., Editor. 2013.
18. Team;, R.C., *R: A language and envrionment for statistical computing. R Foundation for Statistical Computing.* 2012: Vienna, Austria.
19. Wellenius, G.A., et al., *Particulate Air Pollution and the Rate of Hospitalization for Congestive Heart Failure among Medicare Beneficiaries in Pittsburgh, Pennsylvania.* American Journal of Epidemiology, 2005. **161**(11): p. 1030-1036.

20. Casey, J.A., et al., *Using Electronic Health Records for Population Health Research: A Review of Methods and Applications.* Annual Review of Public Health, 2016. **37**(1): p. 61-81.
21. Wolsk, E., et al., *Contribution of cardiac and extra-cardiac disease burden to risk of cardiovascular outcomes varies by ejection fraction in heart failure.* (1879-0844 (Electronic)).

Figure 5a.1. Associations (OR and 95% CI) of quartiles of UNGD pad activity metric with hospitalization in heart failure phenotype strata. The reference group is the HF*p*EF subjects in the first quartile of the pad metric.
Figure 5a.2. Associations (OR and 95% CI) of quartiles of UNGD spud activity metric with hospitalization in heart failure phenotype strata. The reference group is the HF*p*EF subjects in the first quartile of the spud metric.
Figure 5a.3. Associations (OR and 95% CI) of quartiles of UNGD stim activity metric with hospitalization in heart failure phenotype strata. The reference group is the HF*p*EF subjects in the first quartile of the stimulation metric.
Figure 5a.4. Associations (OR and 95% CI) of quartiles of UNGD production activity metric with hospitalization in heart failure phenotype strata. The reference group is the HF*p*EF subjects in the first quartile of the production metric.
Figure 5a.5. Associations (OR and 95% CI) of quartiles of UNGD pad activity metric with hospitalization in heart failure severity strata. The “More Severe Heart Failure” group includes HF*p*EF and HF*r*EF and the “Less Severe Heart Failure” group includes eMERGE not applied and eMERGE no phenotype. The reference group is the less severe heart failure subjects in the first quartile of the pad metric. P-values were obtained from chi^2 tests of the three cross-product terms between the severity indicator and each pad metric quartile.
Figure 5a.6. Associations (OR and 95% CI) of quartiles of UNGD spud activity metric with hospitalization in heart failure phenotype strata. The “More Severe Heart Failure” group includes HF*p*EF and HF*r*EF and the “Less Severe Heart Failure” group includes eMERGE not applied and eMERGE no phenotype. The reference group is the less severe heart failure subjects in the first quartile of the spud metric. P-values were obtained from chi^2 tests of the three cross-product terms between the severity indicator and each spud metric quartile.
Figure 5a.7. Associations (OR and 95% CI) of quartiles of UNGD stimulation activity metric with hospitalization in heart failure phenotype strata. The “More Severe Heart Failure” group includes HF*p*EF and HF*r*EF and the “Less Severe Heart Failure” group includes eMERGE not applied and eMERGE no phenotype. The reference group is the less severe heart failure subjects in the first quartile of the stim metric. P-values were obtained from chi^2 tests of the three cross-product terms between the severity indicator and each stimulation metric quartile.
Figure 5a.8. Associations (OR and 95% CI) of quartiles of UNGD production activity metric with hospitalization in heart failure phenotype strata. The “More Severe Heart Failure” group includes HF*p*EF and HF*r*EF and the “Less Severe Heart Failure” group includes eMERGE not applied and eMERGE no phenotype. The reference group is the less severe heart failure subjects in the first quartile of the production metric. P-values were obtained from chi^2 tests of the three cross-product terms between the severity indicator and each production metric quartile.

Table 5a.1. Selected subject characteristics by phenotype status, at time of randomly selected case event or control encounter

Total subjects n = 9054	Not phenotyped n = 5702	HF*r*EF n = 1739	HF*p*EF n = 1613	p-value*
Sex, n (%)				
Male	2978 (52.2)	1107 (63.7)	648 (40.2)	
Female	2724 (47.8)	632 (36.3)	965 (59.8)	< 0.001
Race/ethnicity, n (%)				
White	5527 (96.9)	1705 (98.0)	1583 (98.1)	
Black	95 (1.7)	12 (0.7)	13 (0.8)	
Hispanic	59 (1.0)	16 (0.9)	13 (0.8)	
Other	18 (0.3)	5 (0.3)	4 (0.3)	
Missing	3 (0.05)	1 (0.06)	0 (0.0)	0.05
Age at hospitalization or at control selection date, years, mean (SD)	70.5 (13.0)	70.3 (12.5)	74.0 (11.2)	< 0.0001
Age category at first event, n (%) years				
> 18-30	49 (0.9)	11 (0.6)	1 (0.06)	
> 30-40	73 (1.3)	26 (1.5)	10 (0.6)	
> 40-50	281 (4.9)	80 (4.6)	36 (2.2)	
> 50-60	768 (13.5)	234 (13.6)	152 (9.4)	
> 60-70	1337 (23.5)	402 (23.1)	319 (19.8)	
> 70-80	1596 (28.0)	524 (30.1)	530 (32.9)	
> 80-90	1526 (26.8)	449 (25.8)	536 (33.2)	
> 90-100	72 (1.3)	10 (0.6)	29 (1.8)	< 0.001
Community type, n (%)				
Borough	1735 (30.4)	578 (33.2)	512 (31.7)	
Township	3245 (56.9)	998 (57.4)	931 (57.7)	
Census tract (city)	722 (12.7)	163 (9.4)	170 (10.5)	0.001
Community socioeconomic deprivation (CSD),** SD units, quartiles		338 (19.4) 436 (25.1) 513 (29.5)	312 (19.3) 391 (24.2) 494 (30.6)	0.6

Total subjects n = 9054	Not phenotyped n = 5702	HF*r*EF n = 1739	HF*p*EF n = 1613	p-value*
1	1046 (18.3)	452 (26.0)	416 (25.8)	
2	1415 (24.8)			
3	1672 (29.3)			
4	1569 (27.5)			
Patient status at end of study, n (%)				
Alive	3969 (69.6)	1058 (60.8)	974 (60.4)	
Deceased	1733 (30.4)	681 (39.2)	639 (39.6)	< 0.001
Distance to major road † (meters), mean (SD)	2794 (4228)	2877 (4295)	2603 (4219)	0.2
Distance to minor road † (meters), mean (SD)	1600 (2400)	1536 (2271)	1423 (2049)	0.02
Distance to hospital/clinic (meters), mean (SD)	6914 (8484)	6531 (8112)	6111 (7778)	0.002
Smoking status at event, n (%)				
Current	721 (12.6)	247 (14.2)	150 (9.3)	
Former	2624 (46.0)	856 (49.2)	749 (46.4)	
Never	2357 (41.3)	636 (36.6)	714 (44.3)	< 0.001
Receipt of Medical Assistance‡ n (%)	679 (11.9)	198 (11.4)	172 (10.7)	0.4
Body mass index (BMI) at event, kg/m^2, mean (SD)	31.6 (8.4)	30.4 (7.5)	32.9 (9.0)	< 0.0001
Time since first heart failure diagnosis, days§, mean (SD)	855 (1148)	754 (1099)	670 (1020)	< 0.0001

* p-value obtained from either chi^2 tests (for categorical variables) or analysis of variance (ANOVA) F-test (for continuous variables), comparing events for subjects who had an assigned phenotype (heart failure with reduced ejection fraction [HF*r*EF] or preserved ejection fraction [HF*p*EF]) vs. those who did not have phenotype information

** Community socioeconomic deprivation (CSD) was calculated based on US Census indicators; further information is detailed in the text

† Major & minor roads were identified from the Federal Highway Administration databases; distance from subject's residential address to these roads was

calculated in meters
‡ Medical Assistance, a surrogate for family socioeconomic status, was calculated based on health insurance status at the time of encounters
§ Days from first heart failure diagnosis to the date of case or control event

Table 5a.2. Selected diagnoses and medication use at time of randomly selected case event or control encounter date, by heart failure phenotype status

Total subjects n = 9054	Not phenotyped n = 5702	HF*r*EF n = 1739	HF*p*EF n = 1613	p-value*
Medication use, by class, n (%)**				
Antihypertensive	2227 (39.1)	803 (46.2)	712 (44.1)	< 0.001
Antihyperlipidemic	2467 (43.3)	831 (47.8)	820 (50.8)	< 0.001
Anticoagulant	967 (17.0)	340 (19.6)	367 (22.8)	< 0.001
Chronic obstructive pulmonary disease, n (%)	1268 (22.2)	354 (20.4)	430 (26.7)	< 0.001
Coronary artery disease, n (%)	1182 (20.7)	520 (29.9)	421 (26.1)	< 0.001
Hypertension, n (%)	4284 (75.1)	1331 (76.5)	1372 (85.1)	< 0.001
Myocardial infarction, n (%)	688 (12.1)	340 (19.6)	143 (8.9)	< 0.001
Valve disorder, n (%)	1081 (19.0)	425 (24.4)	481 (29.8)	< 0.001
Diabetes, n (%)	2365 (41.5)	797 (45.8)	783 (48.5)	< 0.001
Chronic kidney disease, n (%)	1479 (25.9)	528 (30.4)	571 (35.4)	< 0.001
Charlson index of morbidity †, mean (SD)	8.3 (3.3)	8.7 (3.4)	9.4 (3.2)	< 0.0001

* p-value obtained from either chi^2 tests (for categorical variables) or analysis of variance (ANOVA) F-test (for continuous variables), comparing events for subjects who had an assigned phenotype (heart failure with reduced ejection fraction [HF*r*EF] or preserved ejection fraction [HF*p*EF]) vs. those who did not have phenotype information

** Relevant medication classes were identified based on the dates of physician orders

† A composite measure of overall morbidity; definition described in text

Table 5a.3. Adjusted associations (odds ratio [OR], 95% confidence interval [CI]) of phenotype indicators with hospitalization, compared to HF*p*EF as reference group

			Main effects of phenotype indicators*
Phenotype Indicator	Indicator value	Number of events** (n)	**OR (95 % CI)**
HF*r*EF***	1	2185	**1.16 (1.00, 1.35)**
	0	9493	Ref
No phenotype	1	1475	**0.62 (0.53, 0.74)**
	0	10203	Ref
eMERGE not applied	1	5988	**0.39 (0.34, 0.44)**
	0	5690	Ref

* Multilevel logistic regression models adjusted for: sex (male vs. female), race/ethnicity (white vs. nonwhite), receipt of Medical Assistance (ever vs. never received prior to case event or control encounter), smoking status (ever vs. never), body mass index (BMI, as a centered and centered squared term), age category at case event or control encounter, days since first heart failure diagnosis code (centered), distance to nearest hospital or clinic (centered), season, region, and year.
*Number of case events or control encounters represented by each phenotype indicator value
***Heart failure with reduced ejection fraction (HF*r*EF)

Table 5a.4. Associations (odds ratio [OR], 95% confidence interval [CI]) of UNGD activity metrics with hospitalization adjusted for heart failure phenotype and other variables in final model*

Metric	Quartile	OR (95 % CI)
Pad preparation	1	Ref
	2	**1.16 (1.01, 1.33)**
	3	**1.52 (1.31, 1.78)**
	4	**1.57 (1.30, 1.89)**
Spud	1	Ref
	2	0.99 (0.83, 1.18)
	3	1.04 (0.86, 1.27)
	4	0.96 (0.77, 1.19)
Stimulation	1	Ref
	2	1.01 (0.82, 1.23)
	3	**1.42 (1.14, 1.77)**
	4	**1.60 (1.26, 2.02)**
Production	1	Ref
	2	0.91 (0.70, 1.17)
	3	1.06 (0.78, 1.45)
	4	**1.55 (1.11, 2.17)**

* Multilevel logistic regression models adjusted for: sex (male vs. female), race/ethnicity (white vs. nonwhite), receipt of Medical Assistance (ever vs. never received prior to case event or control encounter), smoking status (ever vs. never), body mass index (BMI, as a centered and centered squared term), age category at case event or control encounter, days since first heart failure diagnosis code (centered), distance to nearest hospital or clinic (centered), season, region, year, and three phenotype indicators: heart failure with reduced ejection fraction (yes vs. no), eMERGE algorithm applied, but unable to be phenotyped (yes vs. no), and eMERGE algorithm not applied (yes vs. no)

Table 5a.5. Adjusted* associations (odds ratio [OR], 95% confidence interval [CI]) of pad preparation metric, phenotype indicators, and cross-products between the pad preparation metric and phenotype indicators with hospitalization, compared to HF*p*EF as reference group

Metric or indicator	Quartile or indicator value	OR (95 % CI)	p-value for interaction**
Pad preparation	1	Ref	
	2	**0.70 (0.52, 0.95)**	
	3	1.36 (0.98, 1.88)	
	4	**1.56 (1.11, 2.19)**	
HF*r*EF* **	0	Ref	
	1	0.95 (0.70, 1.29)	
No phenotype	0	Ref	
	1	**0.54 (0.38, 0.76)**	
eMERGE not applied	0	Ref	
	1	0.31 (0.24, 0.40)	
Pad metric x HF*r*EF indicator	1	Ref	
	2	**1.74 (1.15, 2.65)**	
	3	1.17 (0.77, 1.79)	
	4	1.02 (0.67, 1.55)	**0.03**
Pad metric x No phenotype	1	Ref	
	2	1.67 (1.05, 2.65)	
	3	0.98 (0.61, 1.56)	
	4	1.01 (0.64, 1.61)	0.05
Pad metric x eMERGE not applied indicator	1	Ref	
	2	**1.92 (1.36, 2.70)**	
	3	1.19 (0.83, 1.69)	
	4	1.00 (0.70, 1.43)	**0.0002**

* Multilevel logistic regression models adjusted for: sex (male vs. female), race/ethnicity (white vs. nonwhite), receipt of Medical Assistance (ever vs. never received prior to case event or control encounter), smoking status (ever vs. never), body mass index (BMI, as a centered and centered squared term), age category at case event or control encounter, days since first heart failure diagnosis code (centered), distance to nearest hospital or clinic (centered), season, region, year, and cross-products between the pad preparation metric and three phenotype indicators: heart failure with reduced ejection fraction (yes vs. no), eMERGE algorithm applied, but unable to be phenotyped (yes vs. no), and eMERGE algorithm not applied (yes vs. no). The reference group for these three indicators was the heart failure with preserved ejection fraction (HF*p*EF) subjects.

**p-value obtained from chi^2 tests for the cross-products between quartiles of UNGD activity and each phenotype indicator variable

***Heart failure with reduced ejection fraction (HF*r*EF)

Table 5a.6. Adjusted* associations (odds ratio [OR], 95% confidence interval [CI]) of the spud metric, phenotype indicators, and cross-products between the spud metric and phenotype indicators with hospitalization, compared to HF*p*EF as reference group

Metric	Quartile or indicator value	OR (95 % CI)	p-value for interaction**
Spud	1	Ref	
	2	0.99 (0.72, 1.37)	
	3	**1.47 (1.05, 2.05)**	
	4	1.09 (0.77, 1.55)	
HF*r*EF* **	0	Ref	
	1	1.28 (0.96, 1.71)	
No phenotype	0	Ref	
	1	**0.72 (0.53, 0.99)**	
eMERGE not applied	0	Ref	
	1	**0.45 (0.35, 0.58)**	
Spud metric x HF*r*EF indicator	1	Ref	
	2	0.89 (0.59, 1.34)	
	3	0.70 (0.46, 1.05)	
	4	1.09 (0.72, 1.65)	0.2
Spud metric x No phenotype	1	Ref	
	2	1.11 (0.71, 1.74)	
	3	0.69 (0.44, 1.09)	
	4	0.71 (0.45, 1.12)	0.1
Spud metric x eMERGE not applied indicator	1	Ref	
	2	1.01 (0.72, 1.42)	
	3	**0.64 (0.46, 0.91)**	
	4	0.82 (0.58, 1.16)	**0.03**

* Multilevel logistic regression models adjusted for: sex (male vs. female), race/ethnicity (white vs. nonwhite), receipt of Medical Assistance (ever vs. never received prior to case event or control encounter), smoking status (ever vs. never), body mass index (BMI, as a centered and centered squared term), age category at case event or control encounter, days since first heart failure diagnosis code (centered), distance to nearest hospital or clinic (centered), season, region, year, and cross-products between the spud metric and three phenotype indicators: heart failure with reduced ejection fraction (yes vs. no), eMERGE algorithm applied, but unable to be phenotyped (yes vs. no), and eMERGE algorithm not applied (yes vs. no). The reference group for these three indicators was the heart failure with preserved ejection fraction (HF*p*EF) subjects.

** p-value obtained from chi^2 tests for the cross-products between quartiles of UNGD activity and each phenotype indicator variable

***Heart failure with reduced ejection fraction (HF*r*EF)

Table 5a.7. Adjusted* associations (odds ratio [OR], 95% confidence interval [CI]) of the stimulation metric, phenotype indicators, and cross-products between the stimulation metric and phenotype indicators with hospitalization, compared to HF*p*EF as reference group

Metric	Quartile or indicator value	OR (95 % CI)	p-value for interaction**
Stim	1	Ref	
	2	1.01 (0.82, 1.23)	
	3	**1.42 (1.14, 1.77)**	
	4	**1.60 (1.26, 2.02)**	
HF*r*EF* **	0	Ref	
	1	**1.40 (1.05, 1.86)**	
eMERGE applied, but no phenotype	0	Ref	
	1	0.73 (0.54, 1.00)	
eMERGE not applied	0	Ref	
	1	**0.49 (0.38, 0.62)**	
Stim metric x HF*r*EF indicator	1	Ref	
	2	0.87 (0.57, 1.31)	
	3	0.85 (0.57, 1.27)	
	4	0.66 (0.44, 1.00)	0.3
Stim metric x eMERGE applied but no phenotype indicator	1	Ref	
	2	0.79 (0.50, 1.24)	
	3	0.89 (0.57, 1.39)	
	4	0.73 (0.47, 1.13)	0.5
Stim metric x eMERGE not applied indicator	1	Ref	
	2	0.79 (0.56, 1.12)	
	3	0.77 (0.54, 1.08)	
	4	**0.66 (0.47, 0.93)**	0.1

* Multilevel logistic regression models adjusted for: sex (male vs. female), race/ethnicity (white vs. nonwhite), receipt of Medical Assistance (ever vs. never received prior to case event or control encounter), smoking status (ever vs. never), body mass index (BMI, as a centered and centered squared term), age category at case event or control encounter, days since first heart failure diagnosis code (centered), distance to nearest hospital or clinic (centered), season, region, year, and cross-products between the stim metric and three phenotype indicators: heart failure with reduced ejection fraction (yes vs. no), eMERGE algorithm applied, but unable to be phenotyped (yes vs. no), and eMERGE algorithm not applied (yes vs. no). The reference group for these three indicators was the heart failure with preserved ejection fraction (HF*p*EF) subjects.

** p-value obtained from chi^2 tests for the cross-products between quartiles of UNGD activity and each phenotype indicator variable

***Heart failure with reduced ejection fraction (HF*r*EF)

Table 5a.8. Adjusted* associations (odds ratio [OR], 95% confidence interval [CI]) of the production metric, phenotype indicators, and cross-products between the production metric and phenotype indicators with hospitalization, compared to HF*p*EF as reference group

Metric	Quartile or indicator value	OR (95 % CI)	p-value for interaction**
Prod	1	Ref	
	2	1.32 (0.91, 1.92)	
	3	1.20 (0.79, 1.81)	
	4	**1.96 (1.26, 3.04)**	
HF*r*EF* **	0	Ref	
	1	**1.38 (1.04, 1.83)**	
eMERGE applied, but no phenotype	0	Ref	
	1	**0.70 (0.52, 0.95)**	
eMERGE not applied	0	Ref	
	1	**0.50 (0.39, 0.64)**	
Prod metric x HF*r*EF indicator	1	Ref	
	2	0.74 (0.49, 1.12)	
	3	0.94 (0.63, 1.41)	
	4	0.72 (0.47, 1.10)	0.3
Prod metric x eMERGE applied but no phenotype indicator	1	Ref	
	2	0.65 (0.42, 1.01)	
	3	1.12 (0.72, 1.76)	
	4	0.91 (0.58, 1.44)	0.1
Prod metric x eMERGE not applied indicator	1	Ref	
	2	0.60 (0.43, 0.86)	
	3	0.79 (0.56, 1.11)	
	4	0.73 (0.52, 1.03)	**0.04**

* Multilevel logistic regression models adjusted for: sex (male vs. female), race/ethnicity (white vs. nonwhite), receipt of Medical Assistance (ever vs. never received prior to case event or control encounter), smoking status (ever vs. never), body mass index (BMI, as a centered and centered squared term), age category at case event or control encounter, days since first heart failure diagnosis code (centered), distance to nearest hospital or clinic (centered), season, region, year, and cross-products between the production metric and three phenotype indicators: heart failure with reduced ejection fraction (yes vs. no), eMERGE algorithm applied, but unable to be phenotyped (yes vs. no), and eMERGE algorithm not applied (yes vs. no). The reference group for these three indicators was the heart failure with preserved ejection fraction (HF*p*EF) subjects.

**p-value obtained from chi^2 tests for the cross-products between quartiles of UNGD activity and each phenotype indicator variable

***Heart failure with reduced ejection fraction (HF*r*EF)

Table 5a.9. Adjusted* associations (odds ratio [OR], 95% confidence interval [CI]) of UNGD metrics, and cross-products between UNGD metrics and the severity indicator

UNGD activity metric	UNGD activity metric OR (95 % CI)**	UNGD activity metric X Heart failure severity indicator***	p-value for global test of all UNGD metric by severity status terms†
Pad metric			
Quartile 1	Reference	Reference	
Quartile 2	**1.30 (1.11, 1.54)**	**0.71 (0.55, 0.23)**	
Quartile 3	**1.56 (1.30, 1.86)**	0.95 (0.73, 1.23)	
Quartile 4	**1.57 (1.28, 1.94)**	1.00 (0.77, 1.29)	**0.03**
Spud metric			
Quartile 1	Reference	Reference	
Quartile 2	1.02 (0.84, 1.24)	0.93 (0.72, 1.20)	
Quartile 3	0.95 (0.77, 1.18)	**1.30 (1.01, 1.68)**	
Quartile 4	0.86 (0.68, 1.09)	**1.34 (1.03, 1.73)**	**0.009**
Stim metric			
Quartile 1	Reference	Reference	
Quartile 2	0.95 (0.76, 1.18)	1.21 (0.93, 1.56)	
Quartile 3	**1.36 (1.07, 1.73)**	1.17 (0.91, 1.51)	
Quartile 4	**1.52 (1.18, 1.95)**	1.22 (0.95, 1.58)	0.4
Prod metric			
Quartile 1	Reference	Reference	
Quartile 2	0.82 (0.63, 1.07)	**1.40 (1.08, 1.80)**	
Quartile 3	1.02 (0.74, 1.41)	1.18 (0.92, 1.52)	
Quartile 4	**1.51 (1.07, 2.14)**	1.12 (0.88, 1.45)	0.08

* Multilevel logistic regression models adjusted for: sex (male vs. female), race/ethnicity (white vs. nonwhite), receipt of Medical Assistance (ever vs. never received prior to case event or control encounter), smoking status (ever vs. never), body mass index (BMI, as a centered and centered squared term), age category at case event or control encounter, days since first heart failure diagnosis code (centered), distance to nearest hospital or clinic (centered), season, region, year, and interactions between the each respective UNGD metric and a binary indicator for having a phenotype

**Odds ratio reflects the main effect of each UNGD activity metric among the phenotype reference group

***e^{β} (95 % CI) for the cross-product between each UNGD metric and severity status indicator (yes vs. no).

† p-value obtained from chi^2 tests for the cross-products between quartiles of UNGD activity, for each metric, and an indicator variable for phenotype status.

Figure 5a.1.

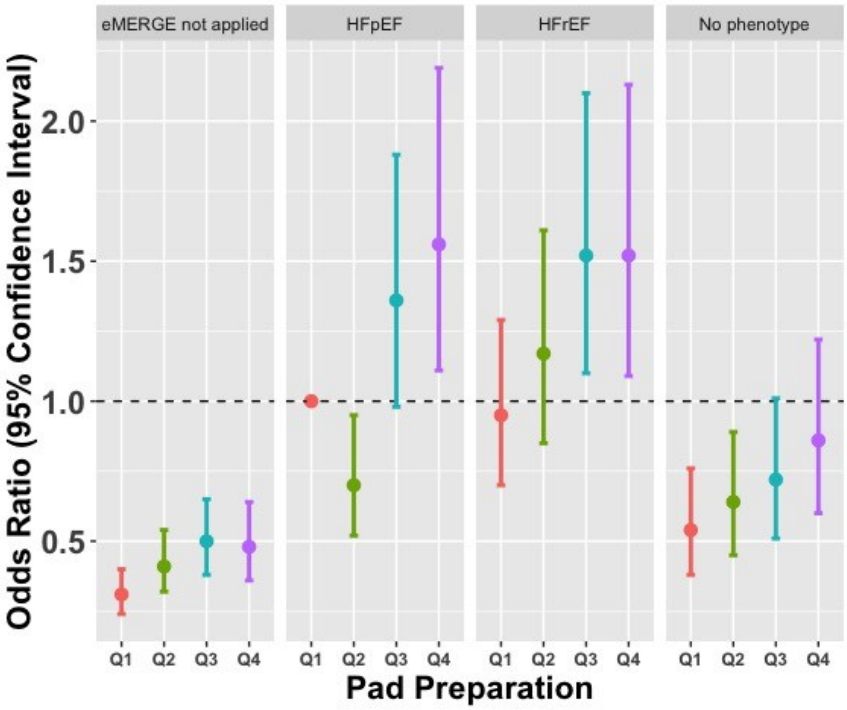


Figure 5a.2.

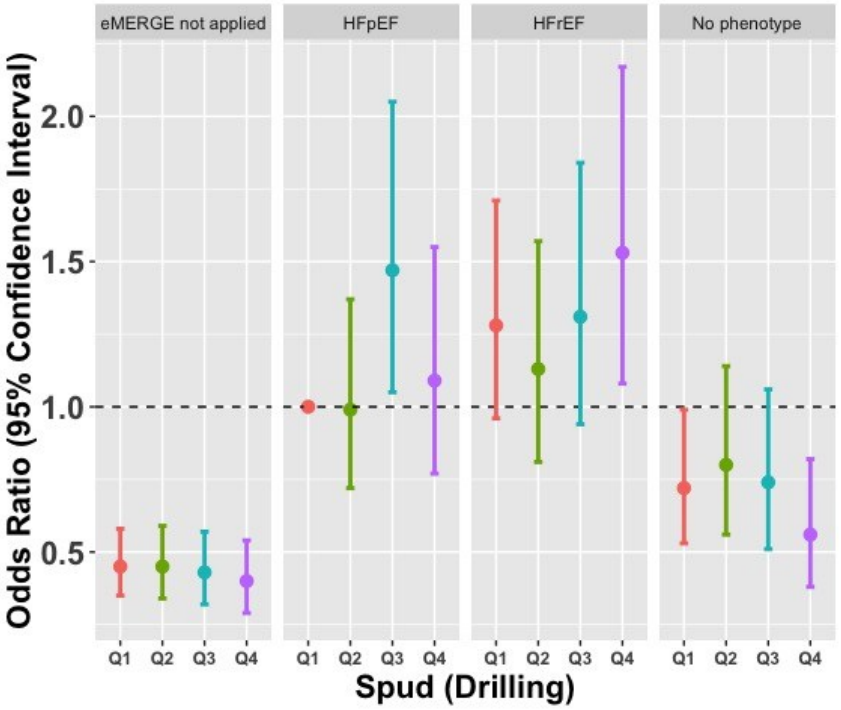
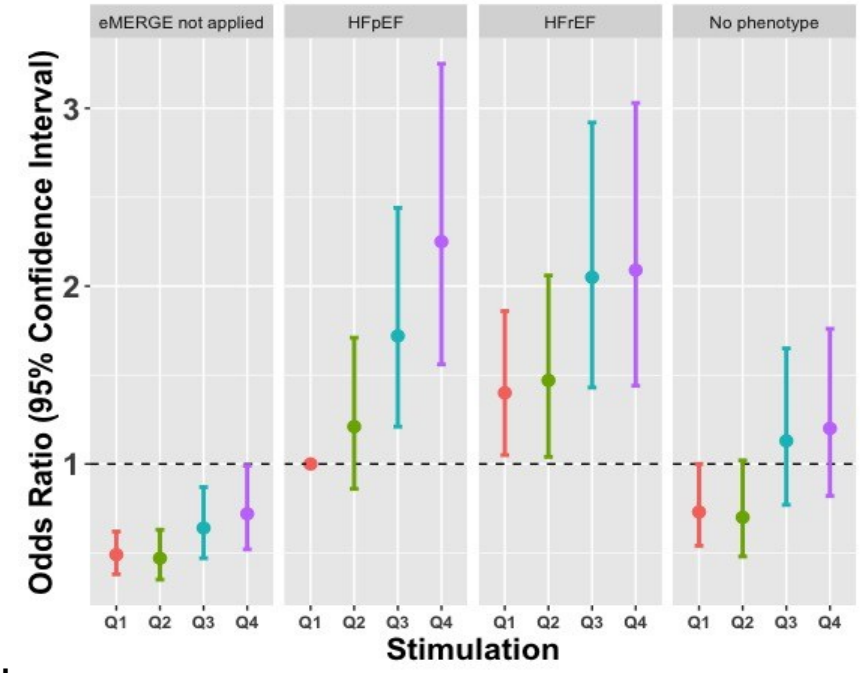


Figure 5a.3



.

Figure 5a.4.

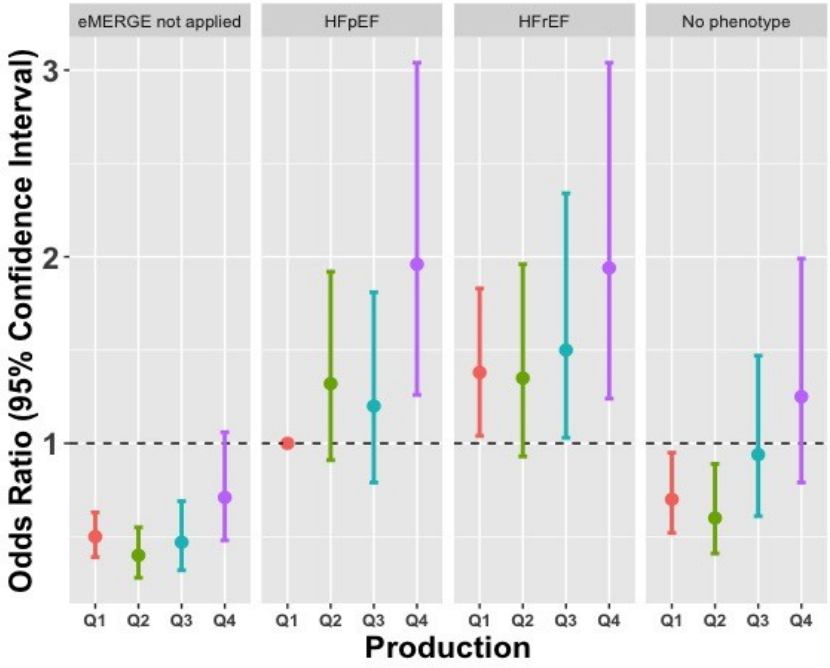


Figure 5a.5.

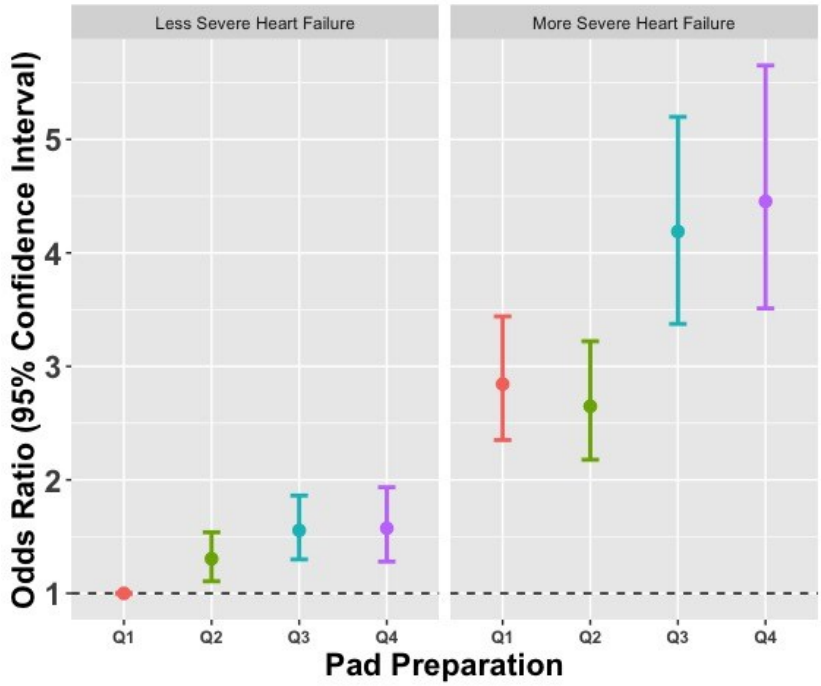


Figure 5a.6.

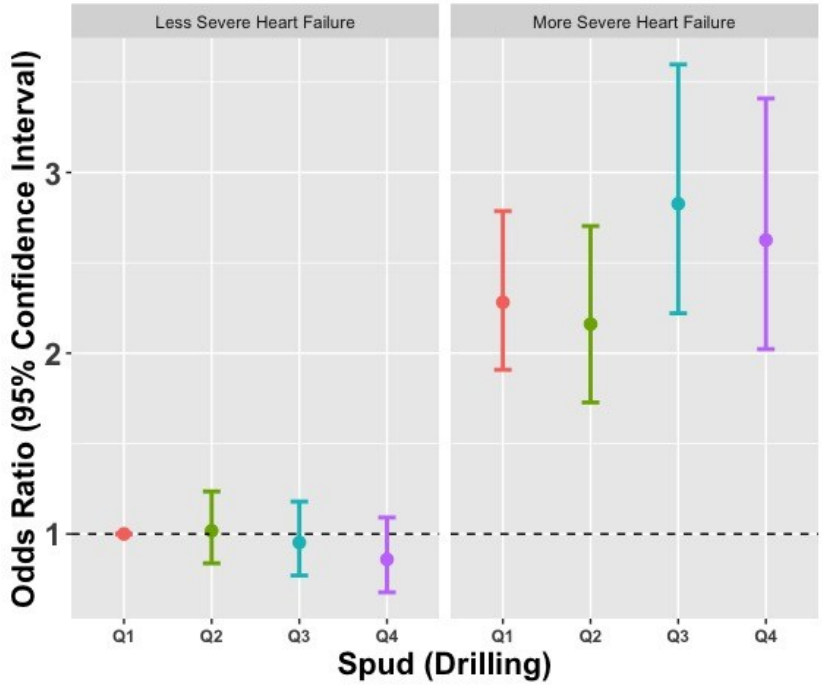


Figure 5a.7.

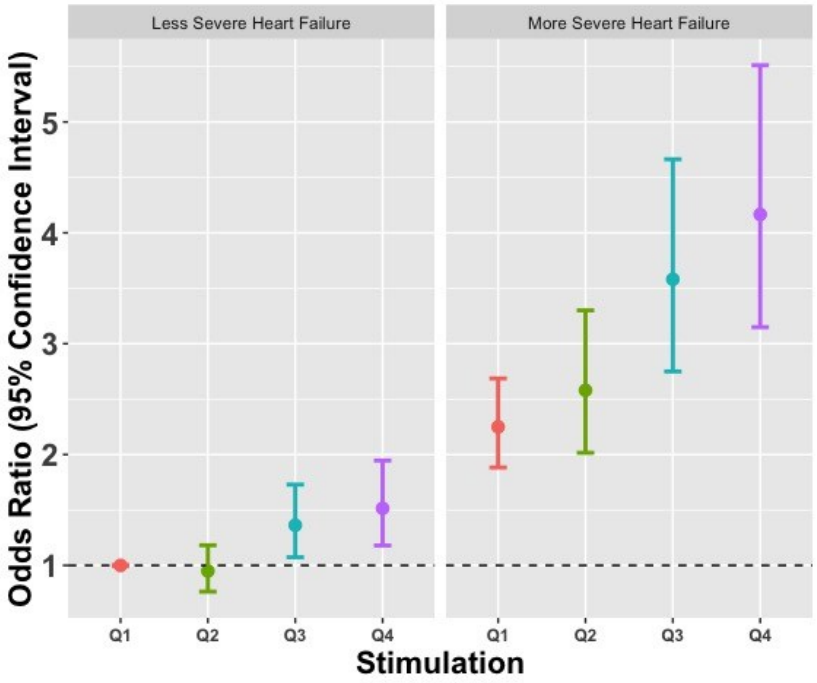
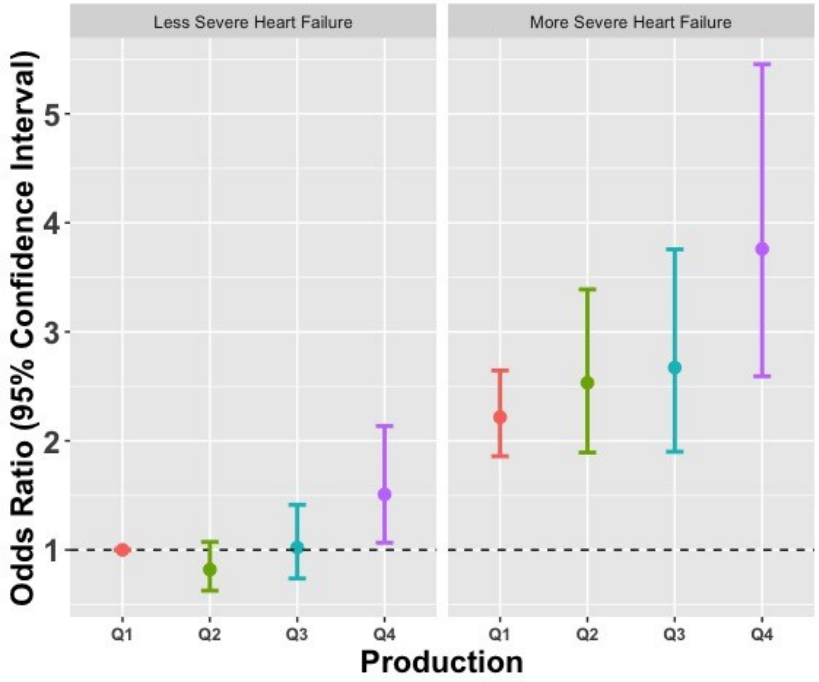


Figure 5a.8.



Chapter 5B: Examination of the association between UNGD activity and B-type natriuretic peptide by heart failure phenotype status

5B1. ABSTRACT

Background: In **Chapter 4**, we observed an exposure-effect association between higher quartiles of UNGD production and greater odds of B-type natriuretic peptide (BNP) levels ≥ 400 pg/mL. In this chapter, given the known pathophysiological differences between heart failure phenotype status (i.e., heart failure with reduced or preserved ejection fraction [HF*r*EF or HF*p*EF]), and because very few large-scale epidemiologic studies of associations between environmental factors and BNP exist, we sought to evaluate if heart failure phenotype modified the association between UNGD activity and BNP levels.

Methods: We used eMERGE heart failure phenotypes that had previously been assigned to 3938 subjects in **Chapter 4**, and we evaluated differences in demographics, comorbidities, and medication use by phenotype categories: not phenotyped (including "eMERGE not applied," and eMERGE applied but not able to phenotype ["eMERGE no phenotype"]), HF*r*EF, and HF*p*EF subjects. We then included cross-products between these phenotype indicators and UNGD metrics in the analysis of effect modification of the UNGD association with greater BNP levels, adjusting for confounding variables. We evaluated the global significance of these cross-product terms using chi^2 tests. We generated linear combinations of stratum-specific odds ratios for each phenotype group and quartile of the four phases of UNGD activity metrics to understand if phenotype category modified the associations observed in **Chapter 4**.

Results: To review, in this analysis and in **Chapter 4** there were 3938 subjects and in **Chapter 5a** there were 9054 subjects, but despite the different samples with the relevant data, we observed some similar associations across chapters in the analysis before evaluation of effect modification. For example, similar to the what we observed in **Chapter 5a**, subjects in this analysis with a HF*p*EF or HF*r*EF phenotype had a greater proportion of comorbidities and medication use compared to subjects without a phenotype, suggesting more severe disease. We also observed that HF*r*EF subjects had higher odds of a BNP level ≥ 400 pg/mL compared to HF*p*EF subjects and to subjects in the no phenotype group; in **Chapter 5a**, HF*r*EF subjects had higher odds of hospitalization compared to HF*p*EF subjects and to subjects in the no phenotype group. Importantly, data suggest that the two samples of persons with BNP data, presented in **Chapters 4** and **5b**, and the subjects in the hospitalization analysis by phenotype (**Chapter 5a**) differed in important ways, with a greater proportion of subjects in this chapter being deceased by the end of the study period compared to the proportion who were deceased in **Chapter 5a**. Unlike differences between phenotyped and not phenotyped subjects in the hospitalization analysis (**Chapter 5a**), phenotyped and not phenotyped subjects in this chapter did not differ by duration of heart failure. In this analysis, after adjusting for phenotype indicators, we observed exposure-effect associations between quartiles of the UNGD production metric and odds of a BNP level ≥ 400 pg/mL, consistent with the findings of **Chapter 4**. We observed, in models of the spud, stimulation, and production metrics, the global test of significance of the cross-products between the indicator for HF*r*EF and quartiles of the spud, stimulation, and production metrics was p = 0.01, 0.001, and 0.0009, respectively; the global test of significance of the cross-product for HF*r*EF and the pad metric yielded a p-value of 0.05. After adjusting for both the main effects of phenotype categories and the cross-products

between phenotype categories and UNGD metrics, we observed an increased odds of BNP levels ≥ 400 pg/mL among HF*p*EF subjects in the fourth quartile of the pad preparation metric (OR [95 % CI] = 1.61 [1.11, 2.35]), and in the third and fourth quartiles of the production metric (OR [95 % CI] = 2.08 [1.33, 3.26], 2.49 [1.54, 2.02], respectively). Although the global test of significance for the cross products between the spud metric and HF*r*EF indicators was p = 0. 01, and p = 0.001 for the stimulation metric, we did not observe exposure-effect associations among the HF*r*EF subjects, and we observed null associations with quartiles of spud and stimulation activity among the HF*p*EF and not phenotyped subjects for these metrics.

Conclusions: This analysis was largely exploratory, and we were unable to determine if HF*r*EF phenotype modified the associations between UNGD activity and BNP levels ≥ 400 pg/mL. Although HF*r*EF was associated with higher BNP levels independent of UNGD activity, our results suggest that HF*p*EF subjects had the strongest association with increasing quartiles of UNGD production and in the 4th quartile of the pad metric, and may therefore be driving the associations we observed between UNGD production metrics and increased odds of BNP ≥ 400 pg/mL in **Chapter 4**. Despite the three significant global tests of cross-products between HF*r*EF and UNGD activity metrics, associations among the HF*r*EF subjects with these UGND metrics did not have a clear exposure-effect trend with increasing quartile, and it is difficult to state whether effect modification was truly present. How associations across **Chapters 4**, **5a**, and **5b** could be influenced by the different samples of persons in each analysis will be discussed.

5B2. INTRODUCTION

B-type natriuretic peptide (BNP), a neuroendocrine hormone, has been increasingly utilized in clinical settings for diagnosis and prognosis of heart failure in persons with the disease [1-3]. BNP levels become elevated when there is an increase

in intracardiac pressure, particularly within the left ventricle [4, 5]. We observed exposure-effect associations between UNGD production activity and odds of a BNP level ≥ 400 pg/mL in **Chapter 4**. A small number of studies have evaluated environmental factors in relation to BNP levels, but these studies are limited by small sample sizes and a lack of distinction between HF*p*EF and HF*r*EF subjects [6, 7].

Despite pathophysiological differences between HF*p*EF and HF*r*EF [8-12], no large scale epidemiologic studies, to date, have evaluated environmental factors in relation to BNP levels among heart failure patients with differing phenotypes (i.e., HF*p*EF vs. HF*r*EF). Given the associations observed in **Chapter 4** between UNGD production activity and odds of BNP levels ≥ 400 pg/mL, we sought to understand if the association between UNGD activity and BNP was modified by phenotype status. Because HF*r*EF subjects generally have higher BNP levels than subjects with HF*p*EF, and because we suspected HF*p*EF subjects might be more susceptible to environmental exposures associated with UNGD activity (e.g., exposure to air pollution, noise, community stress), we did not have a clear *a priori* hypothesis as to which phenotype would have a stronger association between UNGD production and BNP levels, and this analysis was largely exploratory.

5B3. METHODS

We used eMERGE heart failure phenotypes that had been previously assigned [13, 14] to the subjects in our BNP analysis (**Chapter 4**). As first presented in **Chapter 2**, we observed that 2158 of the 3938 subjects in the BNP analysis had the eMERGE algorithm applied, and 1780 of these subjects did not have the eMERGE algorithm applied. Among those who had the eMERGE algorithm applied, 826 were phenotyped as HF*r*EF, 772 were phenotyped as HF*p*EF, and 558 subjects did not have a discernable phenotype. Similar to the application of this phenotyping method to the

hospitalization study subjects (**Chapter 5a**), we determined that missing phenotype information was missing not at random (MNAR), and so we created three binary indicators to categorize subjects into four phenotypic groups: HF*p*EF, HF*r*EF, those without enough information to be phenotyped ("eMERGE not applied"), and those who had enough information to have the eMERGE algorithm applied but did not have a clear HF*p*EF or HF*r*EF phenotype ("eMERGE no phenotype"). For descriptive analysis, we combined the "eMERGE not applied" and "eMERGE no phenotype" groups into a "not phenotyped" group.

We first evaluated frequencies of the number of laboratory orders and the number of subjects categorized into three phenotypic groups: not phenotyped, HF*r*EF, and HF*p*EF to evaluate potential for bias and confounding in our sample. We next evaluated differences in patient characteristics (i.e., demographics, comorbidities, medications) between subjects who were not able to be phenotyped, those who were phenotyped as HF*r*EF, and those who were phenotyped as HF*r*EF. We evaluated differences in these characteristics by analysis of variance (ANOVA) F-tests (for linear variables) and by chi^2 tests (for categorical variables).

In our primary analysis, we used the three binary indicators to reflect our four phenotypic groups (eMERGE not applied, eMERGE no phenotype, HF*r*EF, and HF*p*EF), with HF*p*EF as the reference group as in our analysis for **Chapter 5a**. We added these indicators to the final adjusted GEE models of BNP levels $\geq$ 400 pg/mL (from **Chapter 4**) and assessed the association between each phenotype indicator and odds of a BNP level $\geq$ 400 pg/mL, independent of UNGD activity. Next, we evaluated the main effects of each UNGD activity metric, after adjusting for phenotype groups. Lastly, we evaluated the same GEE models of BNP levels $\geq$ 400 pg/mL but also included cross-product terms between each UNGD metric quartile indicator and the three phenotype group

indicators. We estimated the effects of each indicator, UNGD metric, and interaction term, and we evaluated the global significance of these interaction terms using chi^2 tests. Similar to all evaluations of UNGD activity in this dissertation research, we evaluated each metric of UNGD activity separately. To estimate the odds of BNP levels $\geq$ 400 pg/mL among each phenotype group and for each quartile of UNGD activity, we used the "ggplot2" package in R v.3.4.2 [15] to generate forest plots that visually displayed these odds ratios, with the first quartile of UNGD activity in the HF*p*EF group as the referent. Although Chapter 5a included a *post hoc* analysis of effect modification by an indicator of severity on the associations between UNGD activity and hospitalization, we did not combine HF*r*EF and HF*p*EF groups into an indicator of severity in this analysis because the associations between each phenotype group and BNP, independent of UNGD activity, differed greatly.

5B3. RESULTS

5B.3.1 DESCRIPTIVE CHARACTERISTICS OF PHENOTYPE GROUPS

We observed that the proportion of subjects and the proportion of laboratory orders was relatively similar across phenotype groups (i.e., no phenotype, HF*r*EF, and HF*p*EF), although HF*r*EF and HF*p*EF subjects represented a slightly larger proportion of laboratory orders (22.3% of laboratory orders were from HF*r*EF subjects, and 21.2% of laboratory orders were from HF*p*EF subjects) included in the analysis, compared to the number of subjects each group represented (21.0% of subjects were HF*r*EF and 19.7% of subjects were HF*p*EF, **Table 5b.1**). This suggests that there is little concern for bias or confounding in our analysis due to the distributions of laboratory orders across phenotype groups. Similar to the descriptive results of **Chapter 5a,** we observed some differences in subject characteristics among HF*p*EF, HF*r*EF, and not phenotyped subjects (**Table 5b.2**), with HF*p*EF subjects having a greater mean age and a higher

proportion of females compared to the other two groups ($p < 0.001$ for age and sex). We also observed that a greater proportion of HF*r*EF (52.5%) and HF*p*EF (50.8%) subjects were deceased by the end of the study compared to the not phenotyped subjects (46.7%, $p = 0.008$, **Table 5b.2**). Importantly, we observed some differences compared to the distribution of subjects who were deceased by the end of the study period in **Chapter 5a,** where 30.4% of not phenotyped subjects, 39.2% of HF*r*EF subjects, and 39.6% of HF*p*EF subjects were deceased by the end of the study period ($p < 0.001$, **Table 5a.1**). The subjects in the BNP analysis, and particularly those who had an assigned phenotype, therefore appeared to have more severe disease than the subjects included in the hospitalization analysis. This observation is further supported by differences in mean heart failure duration by phenotype group, which differed in **Chapter 5a** (not phenotyped = 855 days; HF*r*EF = 754 days; HF*p*EF = 670 days, $p < 0.0001$, **Table 5a.1**), but we did not observe differences related to the mean duration of heart failure across all three of these groups (not phenotyped = 806 days; HF*r*EF = 888 days; HF*p*EF = 801 days, $p = 0.1$) in this analysis (**Table 5b.2**).

Although the subjects in this analysis were generally more severe than those included in the hospitalization analysis (i.e., considering duration of heart failure and proportion of subjects who were deceased by the end of the study period), the distribution of comorbidities by phenotype groups was similar to the distributions of comorbidities by phenotype groups observed in **Chapter 5a**. For example, we observed that HF*p*EF subjects also had a greater prevalence of chronic obstructive pulmonary disease, hypertension, diabetes, and chronic kidney disease than HF*r*EF and not phenotyped subjects (**Table 5b.3**). HF*r*EF subjects had a higher proportion of coronary artery disease and of previous myocardial infarctions (**Table 5b.3**). We observed that HF*r*EF and HF*p*EF subjects had a higher prevalence of antihypertensive, antihyperlipidemic, and anticoagulant medications, however none of these differences

were great enough to suggest statistically significant differences by chi^2 test (p = 0.06, 0.2, and 0.2, for each medication class, **Table 5b.3**). The calculated values of the Charlson index of morbidity differed across phenotype groups, with a mean (SD) value of 8.64 (3.0) in the not phenotyped group, 8.79 (3.0) in the HF*r*EF group, and 9.18 (2.6) in the HF*p*EF group (p < 0.001, **Table 5b.3**).

5B.3.2 Associations of phenotype with BNP and confounding of the associations between UNGD activity and BNP by phenotype

In GEE models of BNP, without including any UNGD activity metrics, HF*r*EF subjects had a higher odds of BNP levels $\geq$ 400 pg/mL (OR [95% CI] = 2.24 [1.84, 2.72]) compared to HF*p*EF subjects, after adjusting for comorbidities, laboratory setting, medication use, and subject characteristics (**Table 5b.4**). We observed null associations of eMERGE no phenotype and eMERGE not applied with BNP (OR [95% CI] = 1.03 [0.83, 1.28], 0.97 [0.80, 1.18], respectively, **Table 5b.4**). After adjusting for these phenotype indicators, we did not observe any significant associations between the pad preparation, spud, or stimulation metrics and BNP $\geq$ 400 pg/mL, however we did observe exposure-effect relations among HF*p*EF subjects (i.e., the reference group) with the 2nd, 3rd and 4th quartiles of the production metric and BNP (OR [95% CI] = 1.36 [1.08, 1.73], 1.45 [1.06, 1.97], 1.54 [1.08, 2.21], respectively, compared to the first quartile of production among HF*p*EF subjects. **Table 5b.5**). These effect estimates are very similar to those reported in **Chapter 4**, without adjustment for phenotype indicators, for the production metric (OR [95 % CI] = 1.36 [1.08, 1.71], 1.42 [1.05, 1.93], 1.52 [1.07, 2.17]), suggesting that phenotype did not confound the associations between UNGD production activity and BNP observed in **Chapter 4**.

5B.3.3 Effect modification by phenotype on associations of UNGD activity with BNP

After calculating phenotype-adjusted odds ratios of UNGD activity with BNP, we next evaluated models that also included cross-products between each phenotype indicator and quartile of UNGD activity metrics. In these models, the global test of significance of the cross-products between an indicator for HF*r*EF subjects and quartiles of the pad preparation, spud, stimulation, and production metrics yielded p = 0.05, 0.01, 0.001, and 0.009, respectively (**Tables 5b.6-9**). After inclusion of the cross-products between phenotype indicators and UNGD metrics, we observed an association between UNGD activity and BNP levels within the HF*p*EF group (compared to the first quartile of each metric within the HF*p*EF group) for the 4th quartile of the pad preparation metric (OR [95 % CI] = 1.61 [1.11, 2.35], **Table 5b.6**) and the 3rd and 4th quartiles of the production metric (OR [95 % CI] = 2.08 [1.33, 3.26], 2.49 [1.54, 4.02], respectively, **Table 5b.9**). Within the reference group (HF*p*EF), we did not see any associations of the spud or stimulation metrics with BNP levels (**Tables 5b.7** and **5b.8**).

After accounting for cross-products between UNGD activity metrics and phenotype indicators (i.e., models presented in **Tables 5b.6-9**), we plotted the linear combinations of the main effect of phenotype indicators, the main effect of UNGD activity, and the cross-products between these two indicators to observe stratum-specific estimates of odds of BNP $\geq$ 400 pg/mL for each phenotype group and quartile of UNGD activity (**Figures 5b.1-4**). In plots for each UNGD activity metric, it was evident that the HF*r*EF subjects had higher odds of BNP $\geq$ 400 pg/mL compared to the reference group, HF*p*EF subjects in the first quartile of each UNGD activity metric. Although there was no discernable exposure-effect association for the HF*r*EF subjects in models of the pad or spud metrics (**Figures 5b.1** and **5b.2**), there appeared to be a U-shaped association between increasing quartiles of UNGD activity and odds of BNP $\geq$ 400 pg/mL in the

models of the stimulation and production metrics (**Figures 5b.3** and **5b.4**), however the confidence intervals for each of these quartiles overlapped, so the shape of this association should be interpreted with caution. We did not observe any exposure-effect associations in models of the pad metric (**Figure 5b.1**), spud metric (**Figure 5b.2**), or stimulation metric (**Figure 5b.3**) for subjects without an assigned phenotype (i.e., eMERGE no phenotype and eMERGE not applied compared to the reference group, HF*p*EF subjects in the first quartile of each UNGD activity metric). However, we did see that, even among these less well-characterized phenotype groups, the 4th quartile of the UNGD production metric was associated with greater odds of BNP $\geq$ 400 pg/mL compared to the reference group, HF*p*EF subjects in the first quartile of UNGD production (**Figure 5b.4**). We also observed a 4th quartile association among HF*p*EF subjects for the pad preparation metric compared to the first quartile of the pad preparation metric (**Figure 5b.1**), although the clearest exposure-effect association and trend was observed for the production metric among the HF*p*EF subjects, with odds of BNP $\geq$ 400 pg/mL that increased with the 3rd and 4th quartiles of UNGD production activity compared to HF*p*EF subjects in the first quartile of the production metric (**Figure 5b.4**).

5B4. Discussion

The purpose of this analysis was to assess whether heart failure phenotypes (HF*p*EF and HF*r*EF) modified the associations evaluated in **Chapter 4** between four UNGD activity metrics and odds of BNP $\geq$ 400 pg/mL. Combined effects of phenotype indicators and cross-products between these phenotype indicators and UNGD activity metrics illustrated that, for the pad preparation metric, HF*p*EF subjects had increased odds of BNP $\geq$ 400 pg/mL in the 4th quartile compared to HF*p*EF subjects in the first quartile. Among the HF*p*EF subjects, we observed an exposure-effect association with

increasing quartiles of the production metric and odds of BNP ≥ 400 pg/mL, compared to the HF*p*EF subjects in the first quartile of the production metric. Because we used three different phenotype groups in our analysis to assess effect modification, with HF*p*EF subjects as the reference group, the global tests of significance of the cross-products between phenotype indicators and UNGD activity metrics did not necessarily indicate effect modification after accounting for the main effects of each phenotype with BNP levels. For example, in models of the spud, stimulation, and production metrics, the global test of significance of the cross-products between the indicator for HF*r*EF and quartiles of the spud, stimulation, and production metrics was p = 0.01, 0.001, and 0.0009, respectively; the global test of significance of the cross-product for HF*r*EF and the pad metric yielded a p-value of 0.05. However, we found that, after adjusting for phenotype indicators and cross-products between these indicators and UNGD activity metric quartiles, associations between the spud and stimulation metrics were null for HF*p*EF and not phenotyped subjects, and the HF*r*EF subjects did not evidence a clear exposure-effect association with these metrics. Although the cross-product terms were significant, determining if phenotype modified the associations between UNGD activity and BNP was difficult because, among the HF*r*EF subjects, there was no clear exposure-effect association.

After adjusting for phenotype indicators and cross-product terms between UNGD activity metrics and phenotype indicators, we observed slightly stronger associations with UNGD production and BNP than we did in **Chapter 4**. There was also an emergence of a 4th quartile association between the pad metric and greater odds of BNP ≥ 400 pg/mL among the HF*p*EF subjects, which was not present in our **Chapter 4** analysis. Unlike **Chapter 5a**, where we observed effect modification by a severity indicator (i.e., combined HF*r*EF and HF*p*EF groups) of the associations between the pad

preparation and spud metrics and hospitalization, we did not evaluate effect modification by an indicator of severity. Our reasoning for this was because, comparing the subjects in this analysis to those included in **Chapter 5a**, we did not observe the same differences in duration of disease by phenotype groups as we did in the hospitalization analysis. Additionally, comparing subjects in this analysis to those in **Chapter 5a**, a greater proportion of subjects were deceased in the BNP analysis across all phenotype groups (48.8%) compared to the overall proportion of subjects who were deceased in the hospitalization analysis (33.4%). Therefore, we did not associate the presence of HF*p*EF or HF*r*EF phenotype with severity in the same way as in **Chapter 5a**, because subjects in this analysis, with or without a phenotype, also had severe outcomes and no discernable difference in duration of heart failure. Although the subjects had a greater proportion of subjects who were deceased by the end of the study period compared to **Chapter 5a**, we do not think that introduced bias into results in this chapter because the proportion of comorbidities and medications across phenotype groups (i.e., comparing HF*p*EF, HF*r*EF, and not phenotyped subjects) was similar to the distribution observed across these groups in **Chapter 5a**.

A challenge to the interpretation of the results of this study is that the environmental epidemiology of BNP, and especially the environmental epidemiology of BNP among HF*p*EF and HF*r*EF subjects, is severely limited. We first sought to evaluate the association between UNGD activity and BNP levels in **Chapter 4** because we hypothesized that BNP would be a useful biological marker of the physiological response mechanisms that we suspected UNGD activity would cause (i.e., through air pollution exposure, noise exposure, and experiences of psychosocial stress). Our initial thinking was that all of these exposure pathways would result in systemic inflammation, endothelial dysfunction, and increased systemic blood pressure, which would be correlated with intracardiac pressures and thus BNP levels. Because we observed an

exposure-effect association of UNGD production activity and higher BNP among HF*p*EF subjects and 4th quartile associations with UNGD production activity and higher BNP among the not phenotyped subjects, these findings make biological sense. Considering point estimates only, we observed a U-shaped association between quartiles of UNGD stimulation and production activity and greater BNP levels among the HF*r*EF subjects; however, each quartile had overlapping confidence intervals, so we cannot distinguish the shape of this association from a linear exposure-effect association across quartiles among the HF*r*EF subjects. Because no other studies of this kind exist to which we can compare these results, future studies would need to explore the associations between environmental exposures and BNP levels among HF*r*EF subjects.

A major strength of this study is that we were able to utilize EHR data from subjects who were representative of the general population in Pennsylvania [16] and were able to obtain phenotype information from the application of the validated eMERGE heart failure phenotyping algorithm [13, 14]. We have not been able to identify any other epidemiological studies that utilized the eMERGE heart phenotyping algorithm [13, 14] to examine whether HF*p*EF, HF*r*EF, and no phenotypes modified relations of environmental factors with any heart failure outcomes. Further, this is the first environmental epidemiology study of heart failure phenotypes and BNP. Unlike hospitalization for heart failure (**Chapters 3** and **5a**), BNP as an outcome can be present in asymptomatic patients [3] and can predict mortality in subjects with and without heart failure [17]; and, in this study and in **Chapter 4**, was obtained from both inpatient and outpatient settings. In this sense, the design of this BNP analysis could detect associations between UNGD activity and BNP that occur without the presence of a symptomatic exacerbation of heart failure.

Limitations of this study include the inability to measure dietary, behavioral, and occupational factors from the EHR, limitations that are common to EHR-based

epidemiology studies. Another limitation of this study, similar to a limitation in **Chapter 5a**, is that not all of the subjects in this analysis were able to be assigned a phenotype status from the eMERGE algorithm because the eMERGE algorithm required echocardiogram measures and consistent contact with the health care system (**Appendix A. Heart failure algorithm**). Lastly, a limitation of this study is that we were unable to fully distinguish whether the association between stimulation and production metrics and BNP levels among HF*r*EF subjects was definitively an inverted U-shaped association or a linear exposure-effect association because of the overlapping confidence intervals. However, better characterizing the associations between environmental factors and BNP by heart failure phenotypes is an interesting avenue for future research, and this has been the first study to evaluate these associations.

5B5. CONCLUSIONS

This analysis found that, although BNP levels were independently greater in HF*r*EF subjects compared to HF*p*EF subjects and to subjects without an assigned phenotype, we could not conclude that heart failure phenotype modified associations between UNGD activity and BNP levels $\geq$ 400 pg/mL. However, we did observe an exposure-effect relation with increasing quartiles of UNGD production activity and BNP levels among subjects with HF*p*EF. This supports a growing body of evidence that UNGD activity is associated with adverse health outcomes in exposed populations, and is the first study to evaluate environmental associations with BNP levels among subjects with differing heart failure phenotypes.

5B6. References

1. Santaguida, P.L., et al., *BNP and NT-proBNP as prognostic markers in persons with acute decompensated heart failure: a systematic review.* (1573-7322 (Electronic)).
2. Zhang, C., et al., *Brain Natriuretic Peptide as the long-term cause of mortality in patients with cardiovascular disease: a retrospective cohort study.* International journal of clinical and experimental medicine, 2015. **8**(9): p. 16364-16368.
3. Wang, T.J., et al., *Plasma Natriuretic Peptide Levels and the Risk of Cardiovascular Events and Death.* New England Journal of Medicine, 2004. **350**(7): p. 655-663.
4. Potter, L.R., et al., *Natriuretic peptides: their structures, receptors, physiologic functions and therapeutic applications.* Handbook of experimental pharmacology, 2009(191): p. 341-366.
5. Yancy, C.W., et al., *2017 ACC/AHA/HFSA Focused Update of the 2013 ACCF/AHA Guideline for the Management of Heart Failure.* Journal of the American College of Cardiology, 2017. **70**(6): p. 776.
6. Wellenius, G.A., et al., *Effects of ambient air pollution on functional status in patients with chronic congestive heart failure: a repeated-measures study.* (1476-069X (Electronic)).
7. Vieira, J.L., et al., *Respiratory Filter Reduces the Cardiovascular Effects Associated With Diesel Exhaust Exposure: A Randomized, Prospective, Double-Blind, Controlled Study of Heart Failure: The FILTER-HF Trial.* (2213-1787 (Electronic)).
8. Shah, S.J., D.H. Katz, and R.C. Deo, *Phenotypic Spectrum of Heart Failure with Preserved Ejection Fraction.* Heart failure clinics, 2014. **10**(3): p. 407-418.
9. Tschöpe, C., et al., *Heart failure with preserved ejection fraction: current management and future strategies.* Clinical Research in Cardiology, 2018. **107**(1): p. 1-19.
10. Felker, G.M., et al., *Effect of Natriuretic Peptide–Guided Therapy on Hospitalization or Cardiovascular Mortality in High-Risk Patients With Heart Failure and Reduced Ejection Fraction: A Randomized Clinical TrialNatriuretic Peptide–Guided Therapy for Heart FailureNatriuretic Peptide–Guided Therapy for Heart Failure.* JAMA, 2017. **318**(8): p. 713-720.
11. Harada, E., et al., *B-Type Natriuretic Peptide in Heart Failure With Preserved Ejection Fraction- Relevance to Age-Related Left Ventricular Modeling in Japanese.* (1347-4820 (Electronic)).
12. Kitada, S., et al., *The prognostic value of brain natriuretic peptide in patients with heart failure and left ventricular ejection fraction higher than 60%: a sub-analysis of the J-MELODIC study.* ESC heart failure, 2017. **5**(1): p. 36-45.
13. Bielinski, S.J., et al., *A Robust e-Epidemiology Tool in Phenotyping Heart Failure with Differentiation for Preserved and Reduced Ejection Fraction: the Electronic Medical Records and Genomics (eMERGE) Network.* 2015(1937-5395 (Electronic)).
14. Bielinski;, S.J., *Heart Failure (HF) with Differentiation between Preserved and Reduced Ejection Fraction*, in *PheKB*, M. Clinic., Editor. 2013.
15. Team;, R.C., *R: A language and envrionment for statistical computing. R Foundation for Statistical Computing.* 2012: Vienna, Austria.
16. Casey, J.A., et al., *Using Electronic Health Records for Population Health Research: A Review of Methods and Applications.* Annual Review of Public Health, 2016. **37**(1): p. 61-81.

17. York, M.K., et al., *B-Type Natriuretic Peptide Levels and Mortality in Patients With and Without Heart Failure.* Journal of the American College of Cardiology, 2018. **71**(19): p. 2079-2088.

Table 5b.1. Frequencies of laboratory orders for B-type natriuretic peptide (BNP) and subjects in this analysis by assigned phenotype status

Phenotype	Number of laboratory orders, n (%)	Number of subjects, n (%)
Not phenotyped	3668 (56.4)	2338 (59.4)
HF*r*EF	1452 (22.3)	826 (21.0)
HF*p*EF	1381 (21.2)	774 (19.7)
Total	**6501 (100)**	**3938 (100)**

Table 5b.2. Selected subject characteristics by phenotype status, at time of subjects' first selected lab date

Total subjects n = 3938	**Not phenotyped n = 2338**	**HF*r*EF n = 826**	**HF*p*EF n = 774**	**p-value***
Sex, n (%)				
Male	1223 (52.3)	551 (66.7)	323 (41.7)	
Female	1115 (47.7)	275 (33.3)	451 (58.3)	< 0.001
Race/ethnicity, n (%)				
White	2268 (97.0)	807 (97.7)	759 (98.1)	
Black	36 (1.5)	5 (0.6)	4 (0.5)	
Hispanic	27 (1.2)	10 (1.2)	10 (1.3)	
Other	6 (0.3)	4 (0.5)	1 (0.1)	
Missing	1 (0.04)	0 (0.0)	0 (0.0)	0.2
Age at laboratory date, years, mean (SD)	71.4 (12.3)	69.5 (12.3)	73.3 (10.6)	< 0.001
Community type, n (%)				
Borough				
Township	768 (32.8)	273 (33.1)	246 (31.8)	
Census tract	1264 (54.1)	483 (58.5)	449 (58.0)	
(city)	306 (13.1)	70 (8.5)	79 (10.2)	0.003
Community socioeconomic deprivation (CSD),** SD units, quartiles				
1	579 (24.8)	202 (24.5)	214 (27.6)	
2	588 (25.2)	212 (25.7)	189 (24.4)	
3	568 (24.3)	209 (25.3)	186 (24.0)	
4	603 (25.8)	203 (24.6)	185 (23.9)	0.7
Patient status at end of study, n (%)				
Alive	1244 (53.2)	392 (47.5)	381 (49.2)	
Deceased	1094 (46.7)	434 (52.5)	393 (50.8)	0.008
Distance to major road (meters), mean (SD)	2543 (3947)	3107 (4551)	2433 (4054)	0.001

Total subjects n = 3938	**Not phenotyped n = 2338**	**HF*r*EF n = 826**	**HF*p*EF n = 774**	**p-value***
Distance to minor road (meters), mean (SD)	1514 (2350)	1703 (2476)	1385 (2035)	0.02
Distance to hospital/clinic (meters), mean (SD)	6550 (8053)	6895 (8030)	5960 (7831)	0.06
Smoking status at event, n (%)				
Ever	1399 (59.8)	521 (63.1)	431 (55.7)	
Never	939 (40.2)	305 (36.9)	343 (44.3)	0.01
Receipt of Medical Assistance†, n (%)	228 (9.8)	104 (12.6)	76 (9.8)	0.06
Body mass index (BMI) at event, kg/m^2, mean (SD)	32.1 (7.5)	31.5 (6.9)	33.0 (7.6)	0.000
Time since first HF diagnosis, days‡, mean (SD)	806 (1053)	888 (1106)	801 (1032)	0.1

* p-value obtained from either chi^2 tests (for categorical variables) or ANOVA F-test (for continuous variables), comparing selected characteristics by phenotype group (not phenotyped, heart failure with reduced ejection fraction [HF*r*EF], and heart failure with preserved ejection fraction [HF*p*EF])

** Community socioeconomic deprivation (CSD) was calculated based on US Census indicators; further information is detailed in the text

*** Major & minor roads were identified from the Federal Highway Administration databases; distance from subject's residential address to these roads was calculated in meters

† Medical Assistance, a surrogate for family socioeconomic status, was calculated based on health insurance status at the time of lab date

‡ Days from first HF diagnosis to the date of laboratory order

Table 5b.3. Selected diagnoses and medication use at time of subjects' first selected laboratory date, by heart failure phenotype group

Total subjects n = 3938	**Not phenotyped n = 2338**	**HF*r*EF n = 826**	**HF*p*EF n = 774**	**p-value***
Medication use, by class, n (%)**				
Antihypertensive	970 (41.5)	382 (46.3)	333 (43.0)	0.06
Antihyperlipidemic	1069 (45.7)	403 (48.8)	374 (48.3)	0.2
Anticoagulant	483 (20.7)	186 (22.5)	183 (23.6)	0.2
Chronic obstructive pulmonary disease (COPD), n (%)	501 (21.4)	166 (20.1)	182 (23.5)	0.2
Coronary artery disease (CAD), n (%)	395 (16.9)	230 (27.9)	179 (23.1)	< 0.001
Hypertension, n (%)	1457 (62.3)	545 (66.0)	569 (73.5)	< 0.001
Myocardial infarction, n (%)	188 (8.0)	119 (14.4)	53 (6.9)	< 0.001
Valve disorder, n (%)	376 (16.1)	184 (22.3)	191 (24.7)	< 0.001
Diabetes, n (%)	883 (37.8)	327 (39.6)	367 (47.4)	< 0.001
Chronic kidney disease, n (%)	949 (40.6)	366 (44.3)	366 (47.3)	< 0.001
Charlson index of morbidity*** mean (SD)	8.64 (3.0)	8.79 (3.0)	9.18 (2.6)	< 0.001

* p-value obtained from either chi^2 tests (for categorical variables) or ANOVA F-test (for continuous variables), comparing selected characteristics by phenotype group: not phenotyped, heart failure with reduced ejection fraction (HF*r*EF), and heart failure with preserved ejection fraction (HF*p*EF)
** Relevant medication classes were identified based on the dates of physician orders
*** A composite measure of overall morbidity; definition described in text

Table 5b.4. Adjusted associations (odds ratio [OR], 95% confidence interval [CI]) of phenotype indicators with B-type natriuretic peptide (BNP) value ≥ 400 pg/mL (vs. lower), with heart failure with preserved ejection (HF*p*EF) fraction as the referent group

Phenotype Indicator	**Indicator value**	**Number of labs (n)**	**OR* (95 % CI)**
HF*r*EF**	0	5049	Ref
	1	1452	**2.24 (1.84, 2.72)**
eMERGE applied, but no phenotype	0	2833	Ref
	1	3668	1.03 (0.83, 1.28)
eMERGE not applied	0	3904	Ref
	1	2597	0.97 (0.80, 1.18)

*Odds ratios obtained from generalized estimating equations (GEE) with an exchangeable correlation matrix. Model adjusted for: sex (female vs. male), race/ethnicity (nonwhite vs. white), smoking status (ever vs. never), age at time of laboratory date (centered), laboratory setting (inpatient vs. outpatient), receipt of Medical Assistance (ever vs. never prior to laboratory date), Charlson index (centered and centered squared), body mass index BMI (centered and centered squared), diagnosis of chronic kidney disease, antihyperlipidemic and anticoagulant medication, duration of heart failure (date of lab order – the date of first HF diagnosis, included as a centered and centered squared term), region, and year

**Heart failure with reduced ejection fraction (HF*r*EF)

Table 5b.5. Phenotype indicator-adjusted associations (odds ratio [OR], 95% confidence interval [CI]) of unconventional natural gas development (UNGD) with B-type natriuretic peptide (BNP) value ≥ 400 pg/mL (vs. lower)

Metric	**Quartile**	OR* (95 % CI)
Pad preparation	1	Ref
	2	0.97 (0.82, 1.14)
	3	1.11 (0.93, 1.34)
	4	1.07 (0.88, 1.31)
Spud	1	Ref
	2	1.10 (0.89, 1.36)
	3	1.02 (0.79, 1.31)
	4	1.10 (0.83, 1.45)
Stimulation	1	Ref
	2	1.02 (0.86, 1.22)
	3	1.02 (0.81, 1.28)
	4	1.01 (0.79, 1.29)
Production	1	Ref
	2	**1.36 (1.08, 1.73)**
	3	**1.45 (1.06, 1.97)**
	4	**1.54 (1.08, 2.21)**

*Odds ratios obtained from generalized estimating equations (GEE) with an exchangeable correlation matrix. Model adjusted for: sex (female vs. male), race/ethnicity (nonwhite vs. white), smoking status (ever vs. never), age at time of laboratory date (centered), laboratory setting (inpatient vs. outpatient), receipt of Medical Assistance (ever vs. never prior to laboratory date), Charlson index (centered and centered squared), body mass index BMI (centered and centered squared), diagnosis of chronic kidney disease, antihyperlipidemic and anticoagulant medication, duration of heart failure (date of lab order – the date of first heart failure diagnosis, included as a centered and centered squared term), region, year, and an indicator for heart failure with reduced ejection fraction (HF*r*EF), an indicator for having the phenotyping algorithm applied but not having a discernable phenotype ("eMERGE no phenotype"), and an indicator for not having the algorithm applied ("eMERGE not applied"). The overall phenotype reference group is heart failure with preserved ejection fraction (HF*p*EF).

Table 5b.6. Adjusted associations (odds ratio [OR], 95% confidence interval [CI]) of the pad preparation metric and phenotype indicators with B-type natriuretic peptide (BNP) value ≥ 400 pg/mL (vs. lower), with heart failure with preserved ejection (HF*p*EF) fraction as the referent group

Metric	**Quartile or indicator value**	OR* (95 % CI)	p-value for interaction**
Pad preparation	1	Ref	
	2	1.19 (0.83, 1.68)	
	3	1.18 (0.84, 1.66)	
	4	**1.61 (1.11, 2.35)**	
HF*r*EF* **	0	Ref	
	1	1.10 (0.89, 1.36)	
eMERGE applied, but no phenotype	0	Ref	
	1	1.02 (0.86, 1.22)	
eMERGE not applied	0	Ref	
	1	1.19 (0.85, 1.66)	
Pad metric x HF*r*EF indicator	1	Ref	
	2	0.72 (0.45, 1.13)	
	3	0.83 (0.54, 1.27)	
	4	**0.54 (0.34, 0.84)**	0.05
Pad metric x eMERGE applied but no phenotype indicator	1	Ref	
	2	0.92 (0.55, 1.52)	
	3	1.26 (0.77, 2.07)	
	4	0.81 (0.49, 1.35)	0.3
Pad metric x eMERGE not applied indicator	1	Ref	
	2	0.86 (0.55, 1.33)	
	3	0.75 (0.47, 1.17)	
	4	0.71 (0.46, 1.10)	0.4

*Odds ratios obtained from generalized estimating equations (GEE) with an exchangeable correlation matrix. Model adjusted for: sex (female vs. male), race/ethnicity (nonwhite vs. white), smoking status (ever vs. never), age at time of laboratory date (centered), laboratory setting (inpatient vs. outpatient), receipt of Medical Assistance (ever vs. never prior to laboratory date), Charlson index (centered and centered squared), body mass index BMI (centered and centered squared), diagnosis of chronic kidney disease, antihyperlipidemic and anticoagulant medication, duration of heart failure (date of lab order – the date of first heart failure diagnosis, included as a centered and centered squared term), region, year, and cross-products of the pad preparation metric and each of the following: an indicator for heart failure with reduced ejection fraction (HF*r*EF), an indicator for having the phenotyping algorithm applied but not having a discernable phenotype ("eMERGE no phenotype"), and an indicator for not having the algorithm applied ("eMERGE not applied"), with heart failure with preserved ejection fraction (HF*p*EF) as the referent group

**p-value for global significance of cross-product terms obtained from chi^2 test

***Heart failure with reduced ejection fraction (HF*r*EF)

Table 5b.7. Adjusted associations (odds ratio [OR], 95% confidence interval [CI]) of the spud metric and phenotype indicators with B-type natriuretic peptide (BNP) value ≥ 400 pg/mL (vs. lower), with heart failure with preserved ejection (HF*p*EF) fraction as the referent group

Metric	**Quartile or indicator value**	OR (95 % CI)	p-value for interaction**
Spud	1	Ref	
	2	0.86 (0.58, 1.27)	
	3	1.42 (0.97, 2.08)	
	4	1.30 (0.87, 1.94)	
HF*r*EF* **	0	Ref	
	1	**2.30 (1.62, 3.27)**	
eMERGE applied, but no phenotype	0	Ref	
	1	1.12 (0.77, 1.63)	
eMERGE not applied	0	Ref	
	1	1.08 (0.79, 1.49)	
Spud metric x HF*r*EF indicator	1	Ref	
	2	1.53 (0.95, 2.46)	
	3	0.73 (0.47, 1.13)	
	4	0.91 (0.58, 1.42)	**0.01**
Spud metric x eMERGE applied but no phenotype indicator	1	Ref	
	2	1.24 (0.77, 2.01)	
	3	0.70 (0.44, 1.14)	
	4	0.95 (0.57, 1.59)	0.1
Spud metric x eMERGE not applied indicator	1	Ref	
	2	1.04 (0.71, 1.53)	
	3	0.83 (0.54, 1.27)	
	4	0.71 (0.44, 1.14)	0.3

*Odds ratios obtained from generalized estimating equations (GEE) with an exchangeable correlation matrix. Model adjusted for: sex (female vs. male), race/ethnicity (nonwhite vs. white), smoking status (ever vs. never), age at time of laboratory date (centered), laboratory setting (inpatient vs. outpatient), receipt of Medical Assistance (ever vs. never prior to laboratory date), Charlson index (centered and centered squared), body mass index BMI (centered and centered squared), diagnosis of chronic kidney disease, antihyperlipidemic and anticoagulant medication, duration of heart failure (date of lab order – the date of first heart failure diagnosis, included as a centered and centered squared term), region, year, and cross products between the spud metric and each of the following: an indicator for heart failure with reduced ejection fraction (HF*r*EF), an indicator for having the phenotyping algorithm applied but not having a discernable phenotype ("eMERGE no phenotype"), and an indicator for not having the algorithm applied ("eMERGE not applied"), with heart failure with preserved ejection fraction (HF*p*EF) as the referent group

**p-value for global significance of cross-product terms obtained from chi^2 test

***Heart failure with reduced ejection fraction (HF*r*EF)

Table 5b.8. Adjusted associations (odds ratio [OR], 95% confidence interval [CI]) of the stimulation metric and phenotype indicators with B-type natriuretic peptide (BNP) value ≥ 400 pg/mL (vs. lower), with heart failure with preserved ejection (HF*p*EF) fraction as the referent group

Metric	**Quartile or indicator value**	OR* (95 % CI)	p-value for interaction**
Stim	1	Ref	
	2	0.77 (0.53, 1.10)	
	3	1.25 (0.86, 1.82)	
	4	1.25 (0.84, 1.85)	
HF*r*EF* **	0	Ref	
	1	**2.48 (1.76, 3.50)**	
eMERGE applied, but no phenotype	0	Ref	
	1	0.98 (0.67, 1.42)	
eMERGE not applied	0	Ref	
	1	1.11 (0.81, 1.54)	
Stim metric x HF*r*EF indicator	1	Ref	
	2	1.50 (0.96, 2.37)	
	3	0.82 (0.53, 1.27)	
	4	**0.62 (0.40, 0.97)**	**0.001**
Stim metric x eMERGE applied but no phenotype indicator	1	Ref	
	2	1.41 (0.88, 2.27)	
	3	1.01 (0.62, 1.65)	
	4	0.97 (0.59, 1.59)	0.4
Stim metric x eMERGE not applied indicator	1	Ref	
	2	0.99 (0.67, 1.48)	
	3	0.67 (0.43, 1.04)	
	4	0.84 (0.54, 1.31)	0.2

*Odds ratios obtained from generalized estimating equations (GEE) with an exchangeable correlation matrix. Model adjusted for: sex (female vs. male), race/ethnicity (nonwhite vs. white), smoking status (ever vs. never), age at time of laboratory date (centered), laboratory setting (inpatient vs. outpatient), receipt of Medical Assistance (ever vs. never prior to laboratory date), Charlson index (centered and centered squared), body mass index BMI (centered and centered squared), diagnosis of chronic kidney disease, antihyperlipidemic and anticoagulant medication, duration of heart failure (date of lab order – the date of first heart failure diagnosis, included as a centered and centered squared term), region, year, and cross-products between the stim metric and each of the following: an indicator for heart failure with reduced ejection fraction (HF*r*EF), an indicator for having the phenotyping algorithm applied but not having a discernable phenotype ("eMERGE no phenotype"), and an indicator for not having the algorithm applied ("eMERGE not applied"), with heart failure with preserved ejection fraction (HF*p*EF) as the referent group

**p-value for global significance of cross-product terms obtained from chi^2 test

***Heart failure with reduced ejection fraction (HF*r*EF)

Table 5b.9. Adjusted associations (odds ratio [OR], 95% confidence interval [CI]) of the production metric and phenotype indicators with B-type natriuretic peptide (BNP) value ≥ 400 pg/mL (vs. lower), with heart failure with preserved ejection (HF*p*EF) fraction as the referent group

Metric	**Quartile or indicator value**	OR* (95 % CI)	p-value for interaction**
Prod	1	Ref	
	2	1.20 (0.82, 1.74)	
	3	**2.08 (1.33, 3.26)**	
	4	**2.49 (1.54, 4.02)**	
HF*r*EF* **	0	Ref	
	1	**2.89 (2.03, 4.11)**	
eMERGE applied, but no phenotype	0	Ref	
	1	1.21 (0.82, 1.79)	
eMERGE not applied	0	Ref	
	1	1.10 (0.79, 1.53)	
Prod metric x HF*r*EF indicator	1	Ref	
	2	1.18 (0.76, 1.82)	
	3	0.72 (0.45, 1.13)	
	4	**0.49 (0.31, 0.78)**	**0.0009**
Prod metric x eMERGE applied but no phenotype indicator	1	Ref	
	2	1.24 (0.77, 1.99)	
	3	0.63 (0.38, 1.07)	
	4	0.73 (0.44, 1.21)	**0.02**
Prod metric x eMERGE not applied indicator	1	Ref	
	2	0.94 (0.62, 1.41)	
	3	0.92 (0.58, 1.45)	
	4	0.73 (0.47, 1.15)	0.5

*Odds ratios obtained from generalized estimating equations (GEE) with an exchangeable correlation matrix. Model adjusted for: sex (female vs. male), race/ethnicity (nonwhite vs. white), smoking status (ever vs. never), age at time of laboratory date (centered), laboratory setting (inpatient vs. outpatient), receipt of Medical Assistance (ever vs. never prior to laboratory date), Charlson index (centered and centered squared), body mass index BMI (centered and centered squared), diagnosis of chronic kidney disease, antihyperlipidemic and anticoagulant medication, duration of heart failure (date of lab order – the date of first heart failure diagnosis, included as a centered and centered squared term), region, year, and cross-products between the production metric and each of the following: an indicator for heart failure with reduced ejection fraction (HF*r*EF), an indicator for having the phenotyping algorithm applied but not having a discernable phenotype ("eMERGE no phenotype"), and an indicator for not having the algorithm applied ("eMERGE not applied"), with heart failure with preserved ejection fraction (HF*p*EF) as the referent group

** p-value for global significance of cross-product terms obtained from chi^2 test

***Heart failure with reduced ejection fraction (HF*r*EF)

Figure 5b.1. Associations (OR and 95% CI) of quartiles of UNGD pad activity metric with BNP ≥ 400 pg/mL in heart failure phenotype strata. The reference group is the HF*p*EF subjects in the 1st quartile of the pad metric.
Figure 5b.2. Associations (OR and 95% CI) of quartiles of UNGD spud activity metric with BNP ≥ 400 pg/mL in heart failure phenotype strata. The reference group is the HF*p*EF subjects in the 1st quartile of the spud metric.
Figure 5b.3. Associations (OR and 95% CI) of quartiles of UNGD stimulation activity metric with BNP ≥ 400 pg/mL in heart failure phenotype strata. The reference group is the HF*p*EF subjects in the 1st quartile of the stimulation metric.
Figure 5b.4. Associations (OR and 95% CI) of quartiles of UNGD production activity metric with BNP ≥ 400 pg/mL in heart failure phenotype strata. The reference group is the HF*p*EF subjects in the 1st quartile of the production metric.

Figure 5b.1.

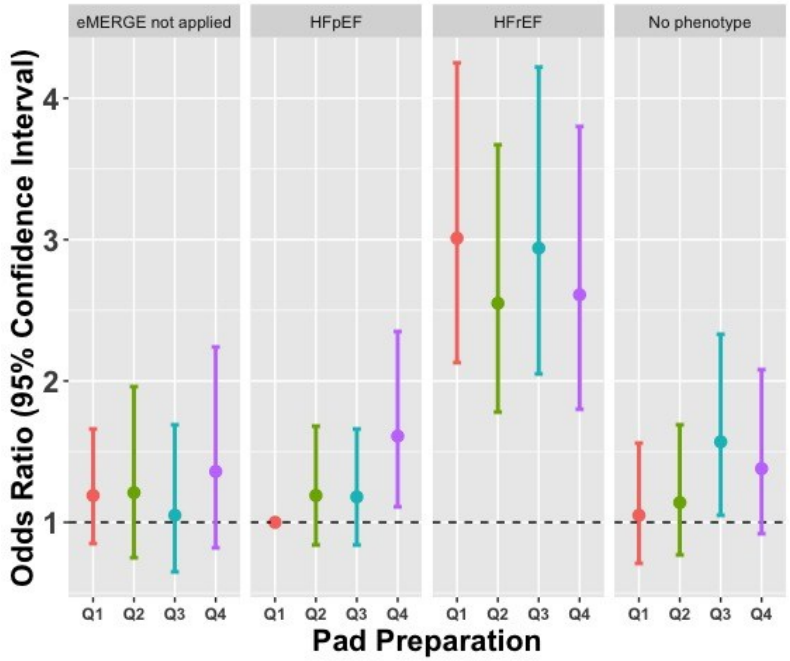


Figure 5b.2.

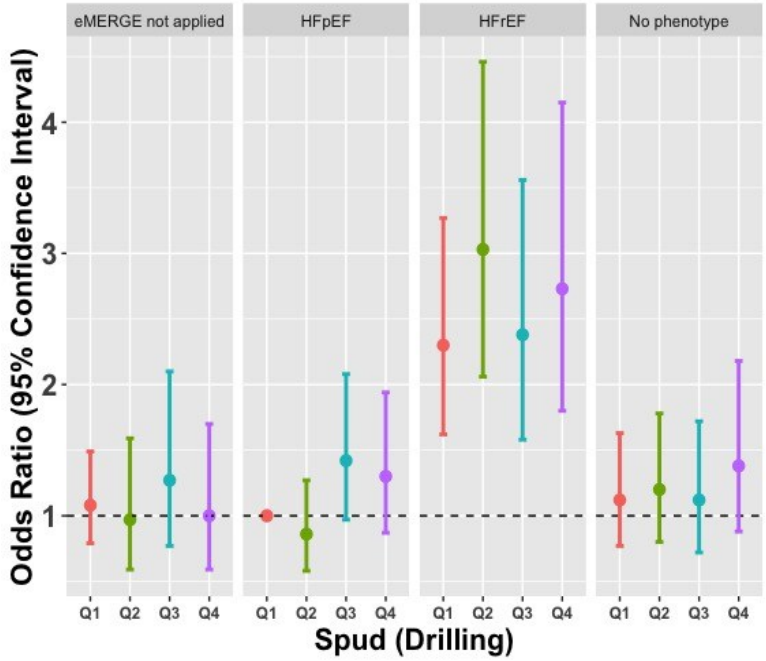


Figure 5b.3.

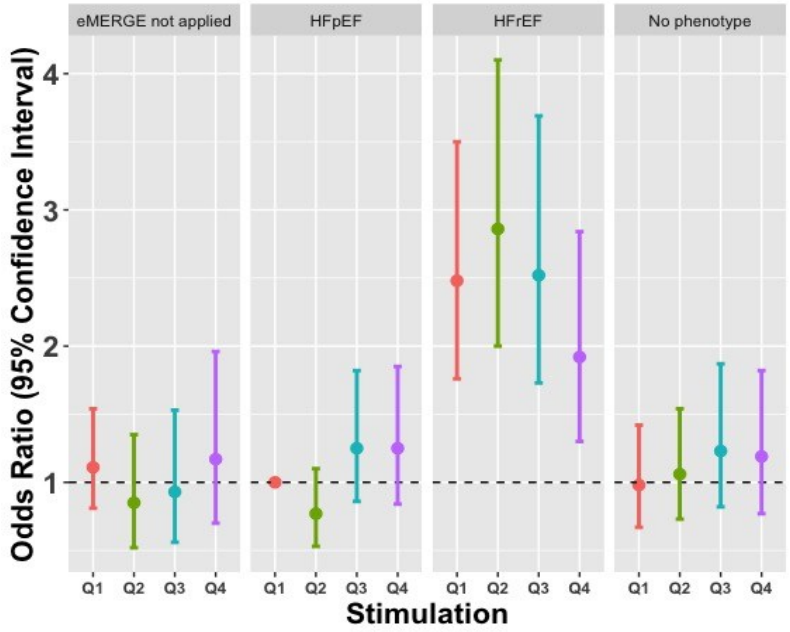
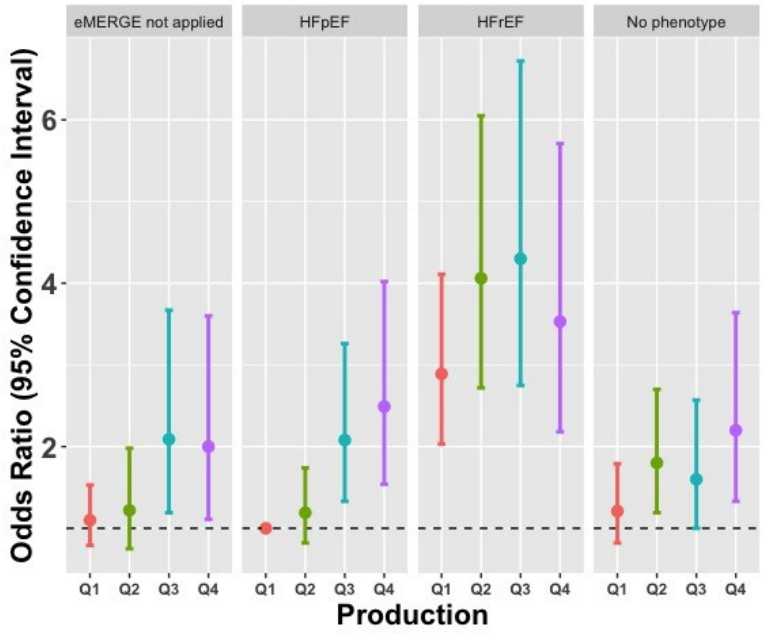


Figure 5b.4.



Chapter 6: Miscellaneous results

This chapter is used for to present results pertinent to the analyses in **Chapters 3** and **4**. Only information that is not presented elsewhere is summarized herein. The analyses are presented in brief form organized by the respective chapters to which they pertain.

6.1 Miscellaneous results pertaining to Chapter 3

- We wanted to know if control subjects who only had one ICD-9 code, or who did not receive any medications for heart failure, could be biasing the results of the case-control study.
- We identified 1106 people were never a case (only served as control) and had no heart failure medication. Of these, 622 subjects had only one icd-9 code for 428.x.
- Descriptive statistics for these 622 individuals are below in **Table 6.1**.

Table 6.1 Descriptive characteristics of control subjects with only one heart failure diagnosis code

Variable	n (%) of 622
Sex, n (%) Male Female	 307 (49.4) 315 (50.6)
Race/ethnicity, n (%) White Black Hispanic Other	 603 (97.0) 11 (1.8) 7 (1.1) 1 (0.2)
Chronic obstructive pulmonary disease, n (%)	144 (23.2)
Coronary artery disease, n (%)	159 (29.6)
Hypertension, n (%)	506 (81.4)
Myocardial infarction, n (%)	70 (11.3)
Valve disorder, n (%)	109 (17.5)
Diabetes, n (%)	260 (41.8)
Chronic kidney disease, n (%)	178 (28.6)

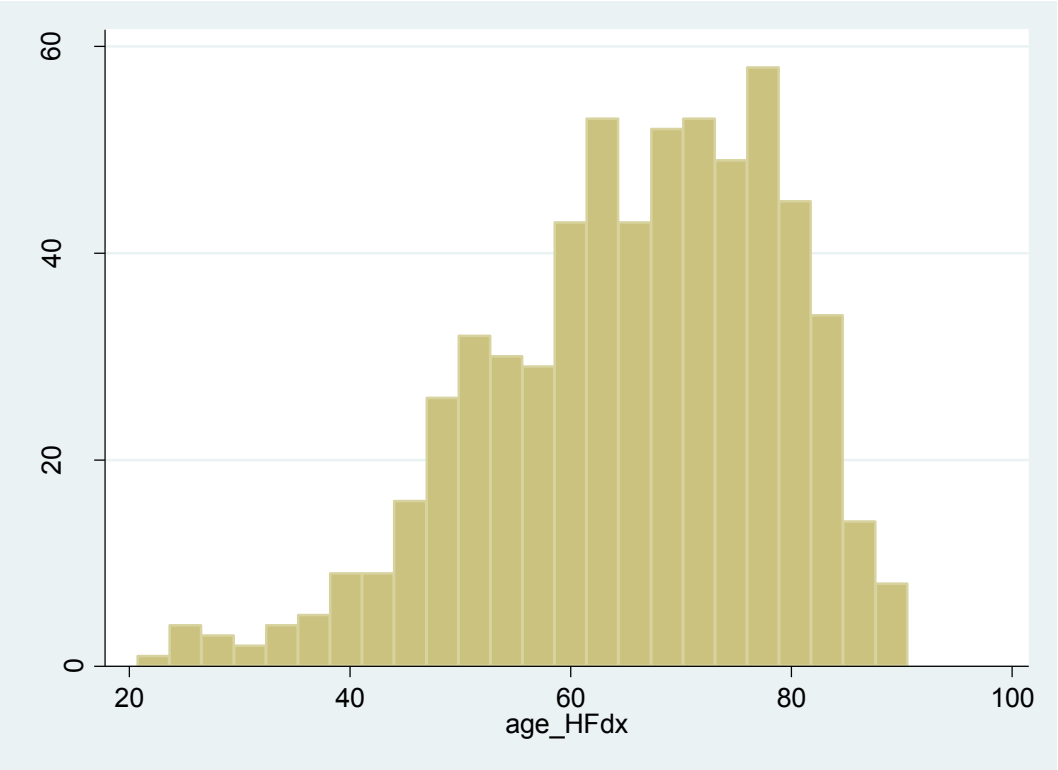


FIGURE 6.4. HISTOGRAM OF AGE AT FIRST HEART FAILURE DIAGNOSIS FOR 622 CONTROLS WITH ONLY ONE HEART FAILURE ICD-9 CODE

- Within the 622 individuals who only had one ICD-9 code for heart failure, we wanted to determine if those without any comorbidities (i.e., hypertension, myocardial infarction, valve disease, coronary artery disease, chronic obstructive pulmonary disease, diabetes) or medications for heart failure (i.e., antihypertensive, antihyperlipidemic, and anticoagulant medications), could be influencing the results in **Chapter 3**.
- We excluded 52 individuals (8.4% of the 622) who had no comorbidities. The age distribution of these individuals is below (**Figure 6.2**).
- We conducted a sensitivity analysis excluding the 52 individuals with no comorbidities, leaving us with 9002 subjects with 11,604 observations.
- Results of this sensitivity analysis (odds ratios and 95% CIs) for odds of hospitalization by quartile of each UNGD metric are displayed in **Table 6.2**.
- We concluded that these 52 individuals were not providing undue influence in our main analysis in **Chapter 3**.

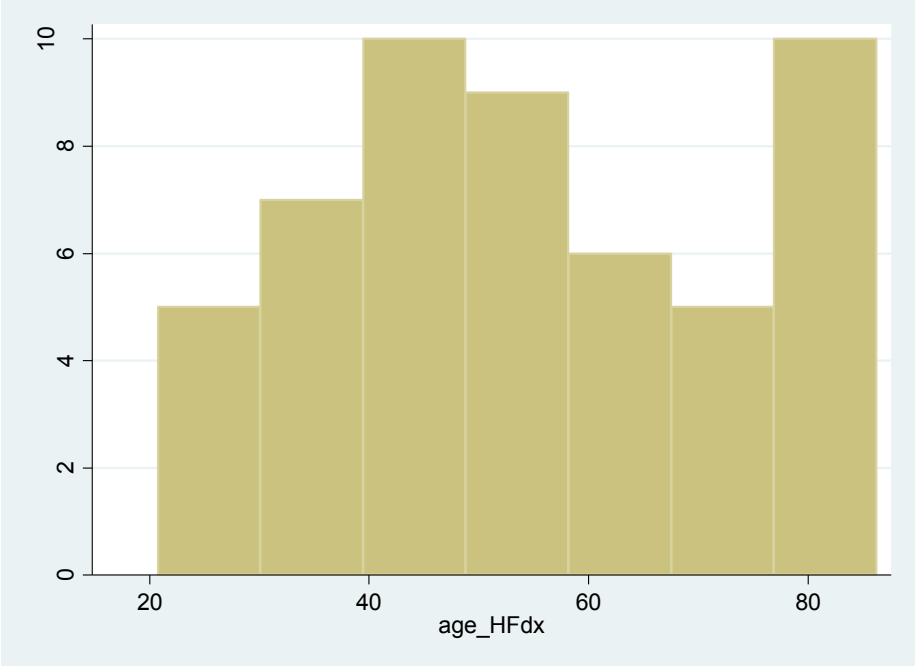


FIGURE 6.5. AGE AT HEART FAILURE DIAGNOSIS FOR THE 52 SUBJECTS WITH ONLY 1 HEART FAILURE DIAGNOSIS, AND NO COMORBIDITIES OR MEDICATIONS

Table 6.2. Adjusted associations of UNGD metrics and odds of hospitalizations

UNGD activity metric	OR (95 % CI)*
Pad metric	
Quartile 1	Reference
Quartile 2	**1.19 (1.01, 1.40)**
Quartile 3	**1.67 (1.37, 2.00)**
Quartile 4	**1.71 (1.36, 2.15)**
Spud metric	
Quartile 1	Reference
Quartile 2	1.01 (0.82, 1.25)
Quartile 3	1.07 (0.84, 1.34)
Quartile 4	0.96 (0.74, 1.25)
Stim metric	
Quartile 1	Reference
Quartile 2	1.01 (0.80, 1.29)
Quartile 3	**1.54 (1.17, 2.01)**
Quartile 4	**1.78 (1.33, 2.38)**
Prod metric	
Quartile 1	Reference
Quartile 2	0.87 (0.64, 1.19)
Quartile 3	1.11 (0.76, 1.62)
Quartile 4	**1.63 (1.07, 2.46)**

*Obtained from multi-level logistic regression models, adjusting for: sex (male vs. female), race/ethnicity (white vs. nonwhite), receipt of Medical Assistance, smoking status (ever vs. never), body mass index (BMI, centered), age matching category, season, observation time (centered), distance to hospital or clinic, year, and region.

6.2 MISCELLANEOUS RESULTS PERTAINING TO CHAPTER 4

- Initially, we wanted to evaluate changes in BNP levels over time in relation to environmental factors. To understand how an individual's trajectory changed over time, we divided the 3938 subjects into 10 groups based on the decile of their first BNP laboratory value. Within each decile, we randomly selected 30 individuals and plotted their BNP measurements over time (**Figures 6.3-6.12**).
- Many subjects only had two BNP values and there was great heterogeneity on how values changed from the first to the second measurement. Most of these values were below typical cutoffs for concern (e.g., 400 pg/mL). The results suggested that regression towards the mean may have been one primary determinant of change, in that low values tended to increase and high values tended to decline.
- Given the limitations of two values for longitudinal analysis and the observation above, we concluded there was little that was likely to be learned by subjecting these values to longitudinal analysis.

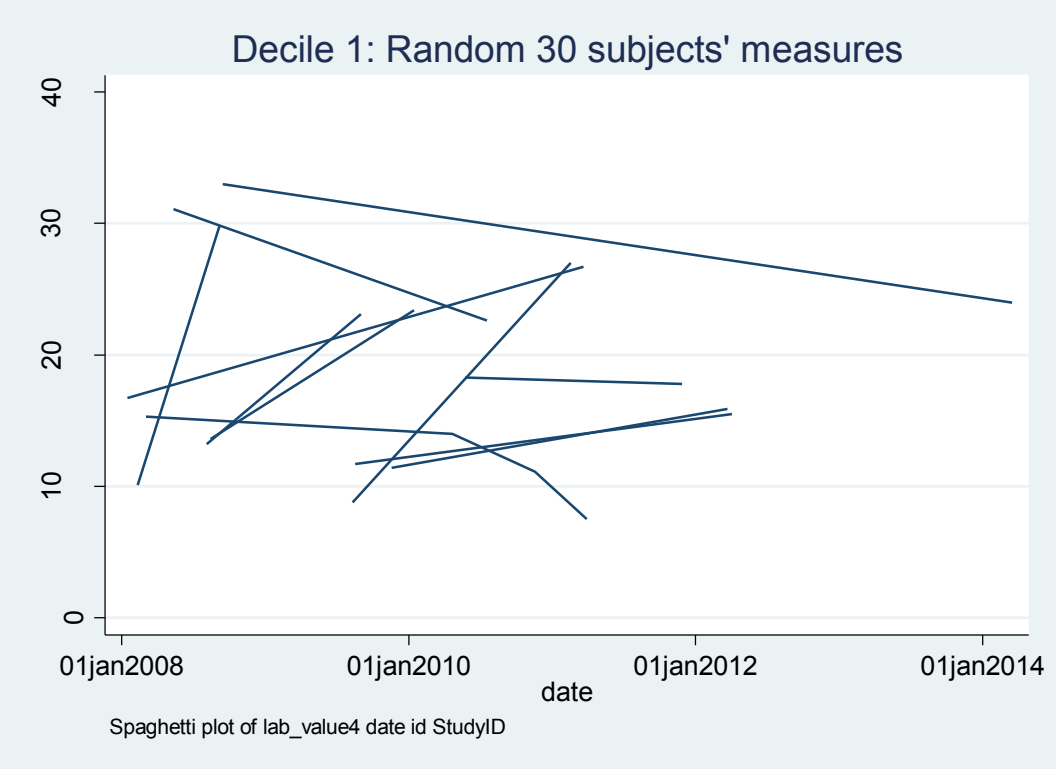


FIGURE 6.6. RANDOM 30 SUBJECTS' BNP VALUES OVER TIME (DECILE 1)

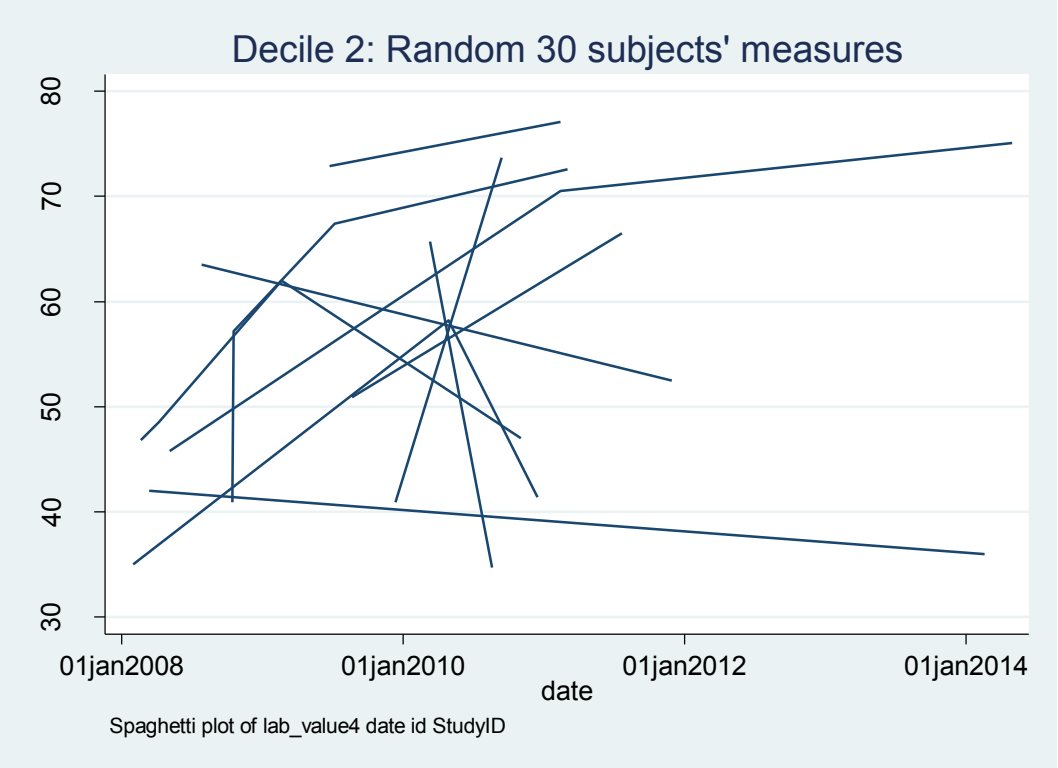


FIGURE 6.7. RANDOM 30 SUBJECTS' BNP VALUES OVER TIME (DECILE 2)

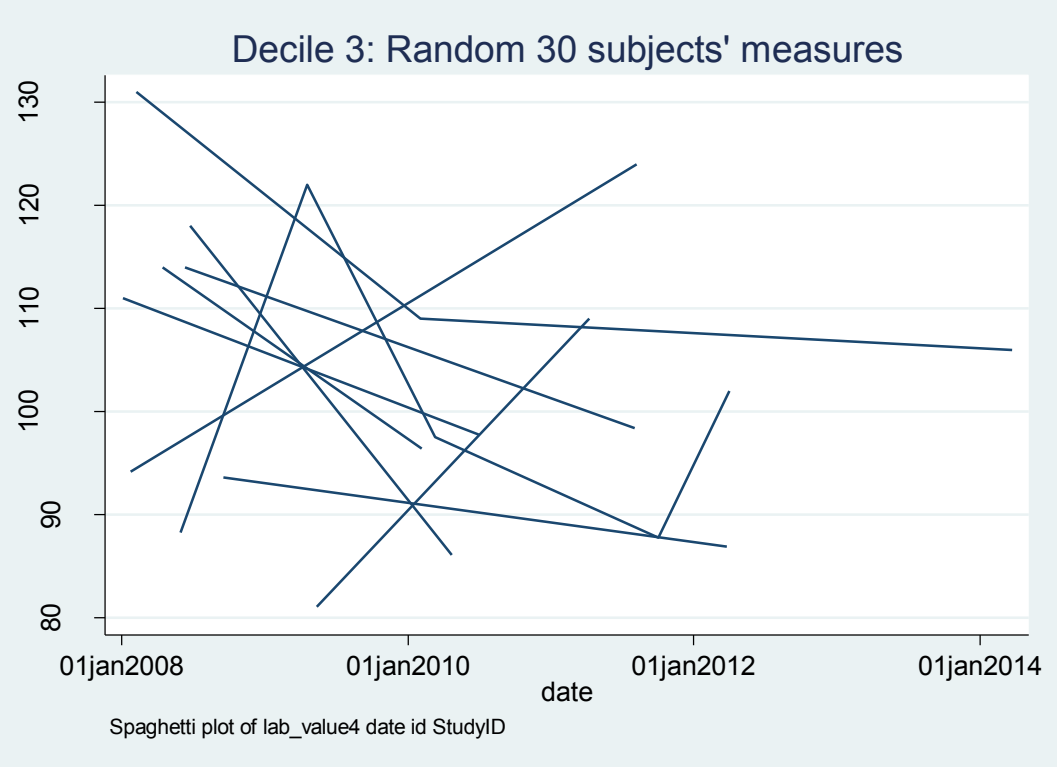


FIGURE 6.8.RANDOM 30 SUBJECTS' BNP VALUES OVER TIME (DECILE 3)

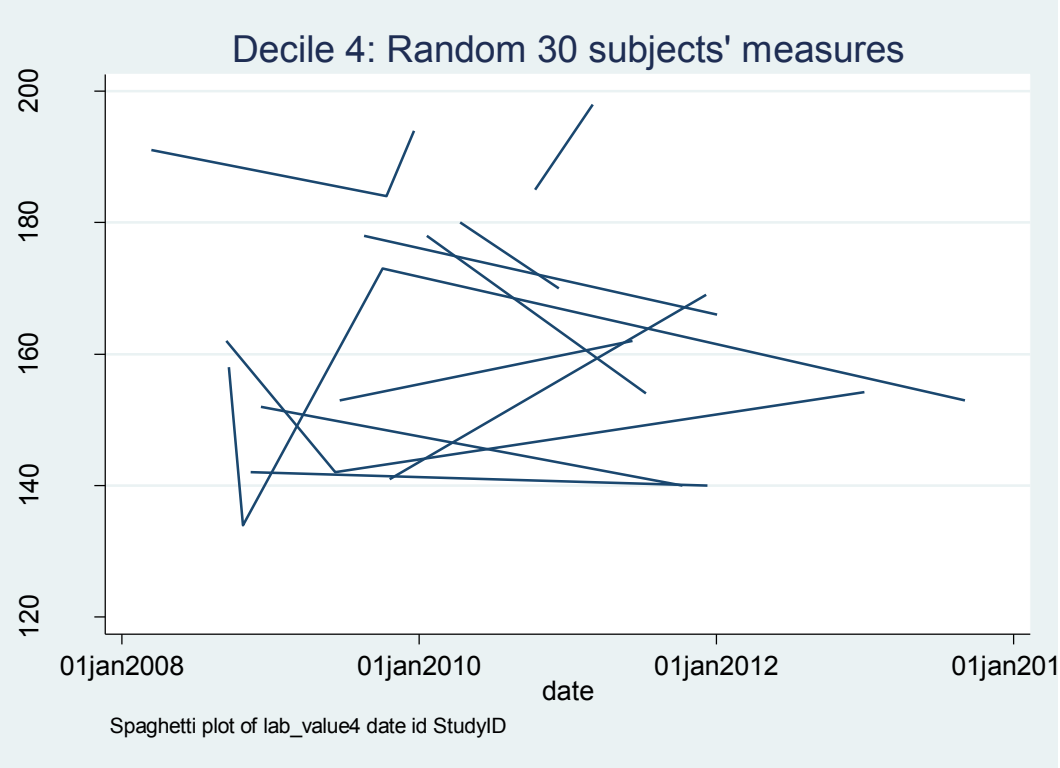


FIGURE 6.9. RANDOM 30 SUBJECTS' BNP VALUES OVER TIME (DECILE 4)

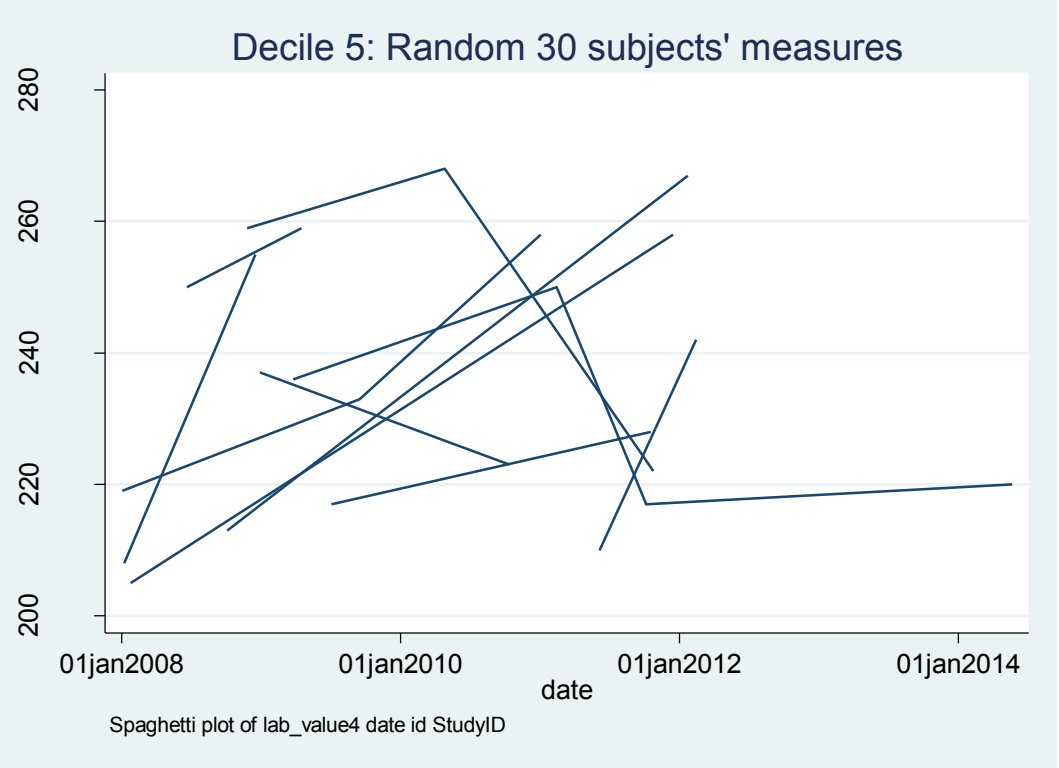


FIGURE 6.10. RANDOM 30 SUBJECTS' BNP VALUES OVER TIME (DECILE 5)

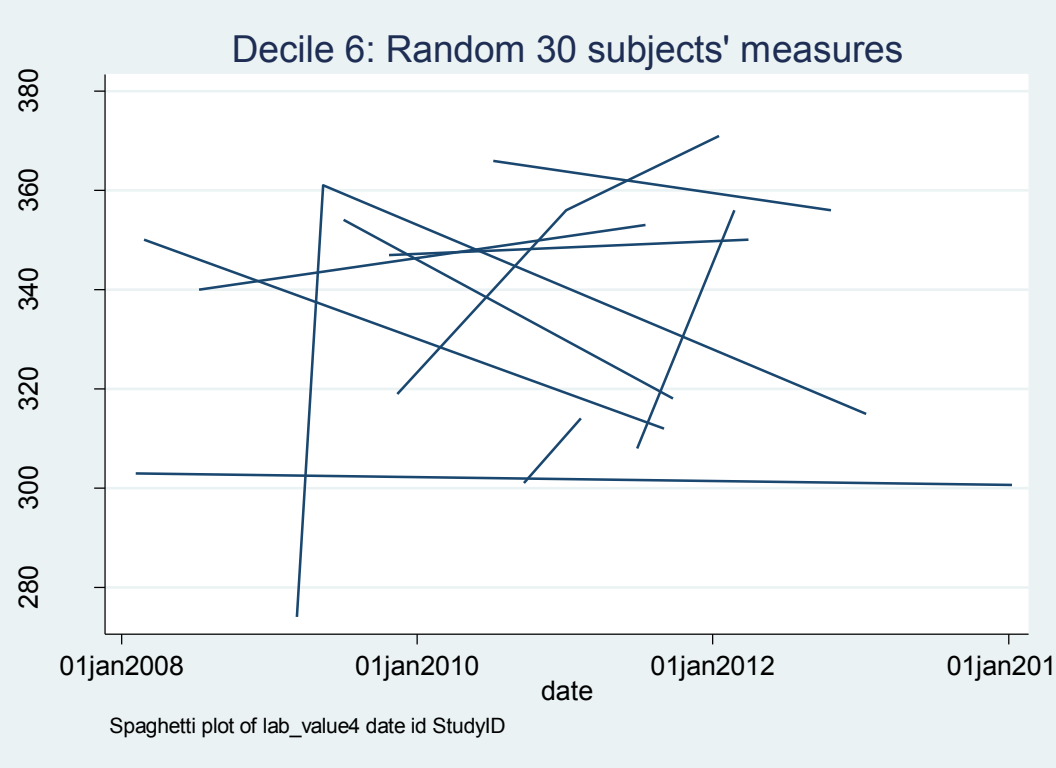


FIGURE 6.11. RANDOM 30 SUBJECTS' BNP VALUES OVER TIME (DECILE 6)

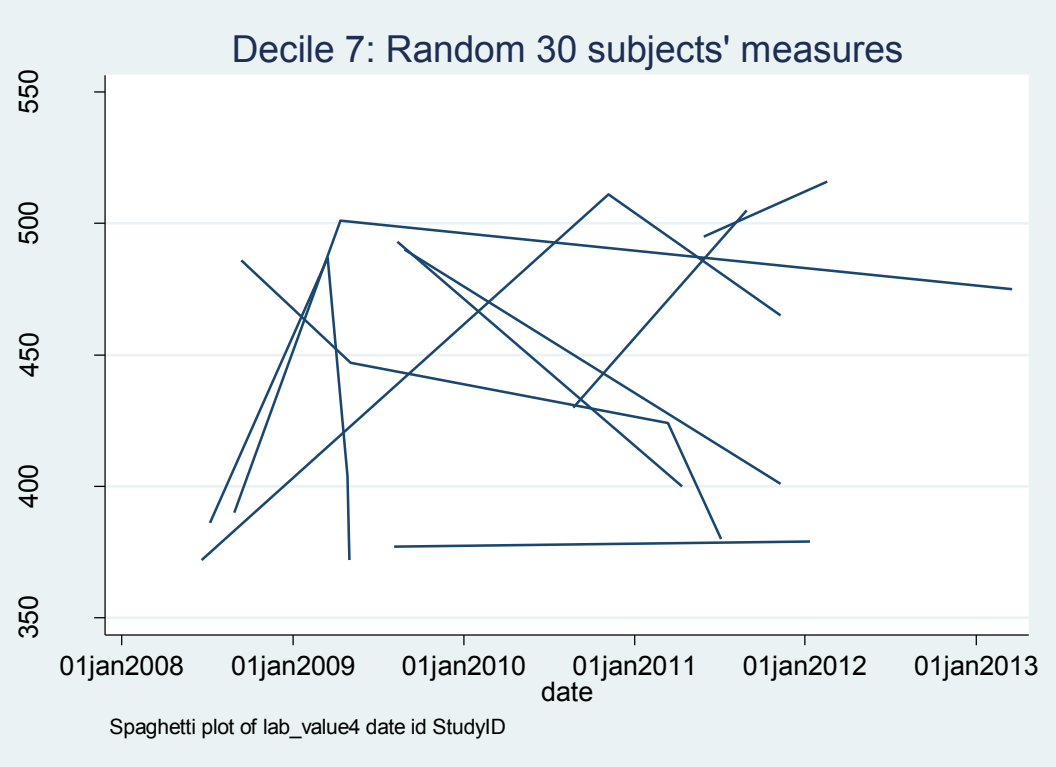


FIGURE 6.12. RANDOM 30 SUBJECTS' BNP VALUES OVER TIME (DECILE 7)

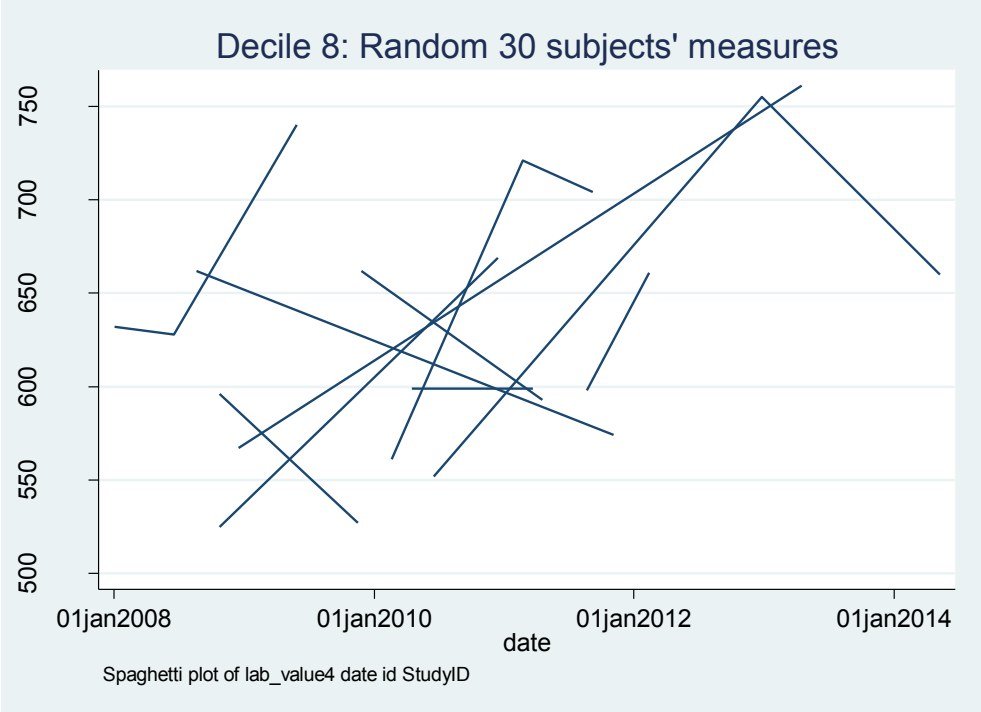


FIGURE 6.13. RANDOM 30 SUBJECTS' BNP VALUES OVER TIME (DECILE 8)

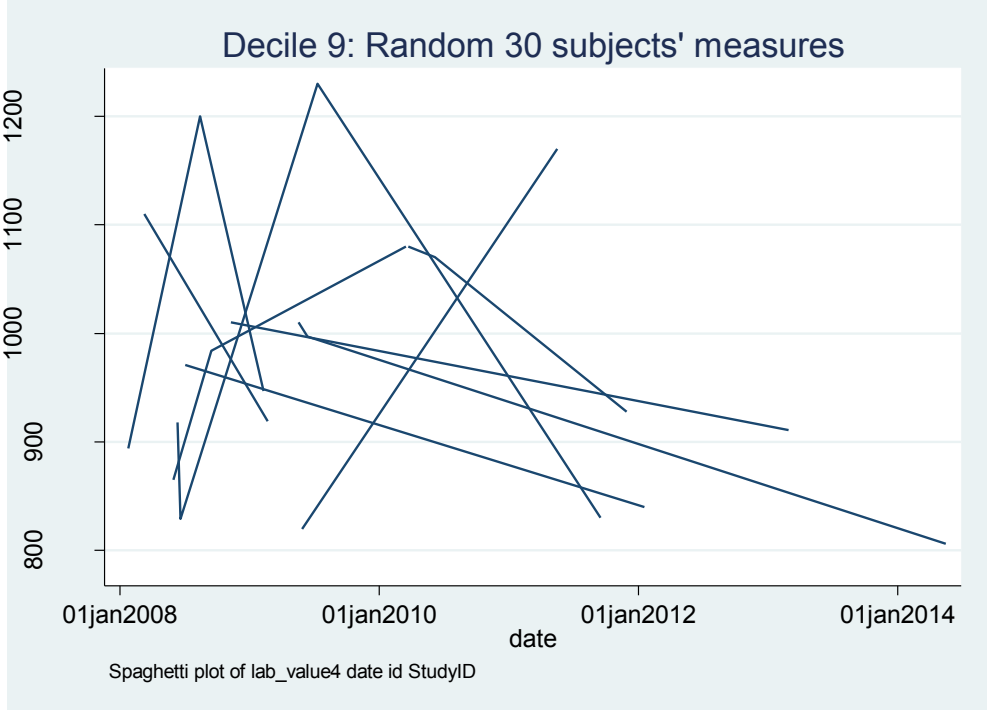


FIGURE 6.14. RANDOM 30 SUBJECTS' BNP VALUES OVER TIME (DECILE 9)

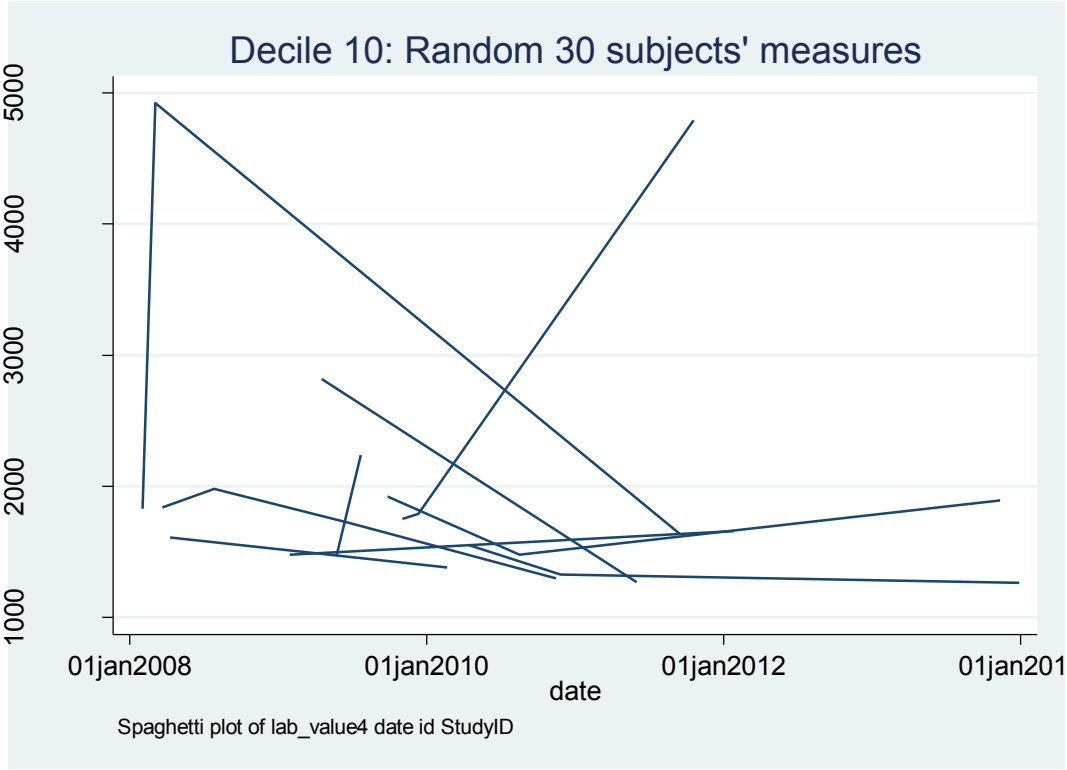


FIGURE 6.15. RANDOM 30 SUBJECTS' BNP VALUES OVER TIME (DECILE 10)

Chapter 7: Discussion

7.1 Summary of findings

This research examined associations between metrics of unconventional natural gas development (UNGD) activity and heart failure outcomes in Pennsylvania. First, we found that among 9054 subjects with a diagnosis of heart failure, 30-day UNGD metrics for the pad preparation, stimulation, and production metrics were associated with greater odds of hospitalization. Second, we found that among 3938 subjects with a diagnosis of heart failure and laboratory measures of B-type natriuretic peptide (BNP), quartiles of the UNGD production activity metric were associated with greater odds of a BNP level $\geq$ 400 pg/mL. Third, we found no evidence consistent with our *a priori* hypothesis that associations of UNGD with hospitalization (and an exploratory analysis of the same effect modification for BNP) were stronger in HF*p*EF (vs. HF*r*EF). Fourth, in a *post hoc* analysis, we found that subjects who were phenotyped as HF*p*EF or HF*r*EF (vs. no phenotype), which we interpreted as a surrogate for more severe heart failure, had stronger associations of UNGD pad preparation, spud, and production metrics with hospitalization (although the global test of the three cross-products was of borderline significance for production). Lastly, in our analysis to evaluate effect modification by heart failure phenotypes on relations of UNGD metrics with BNP levels divided at $\geq$ 400 pg/mL, we found significant or borderline significant global tests for both UNGD pad preparation and production metrics comparing HF*r*EF to HF*p*EF, but the UNGD associations were difficult to interpret given the non-monotonic associations of UNGD pad and production with BNP levels among those with HF*r*EF. phenotype (vs.HF*p*EF subjects).

All of these findings are the first of their kind in both the literature on the environmental epidemiology of UNGD as well as in the literature on the environmental epidemiology of heart failure, its biomarkers, and its phenotypes. Below I discuss the importance and implications of these findings.

7.1.2 Description of goals of this dissertation

The goals of this dissertation research were to:

SA1. Evaluate associations between metrics of UNGD activity and odds of hospitalization in a case-control study.

SA2. Evaluate associations between metrics of UNGD activity and odds of BNP levels $\geq$ 400 pg/mL.

SA3. Evaluate whether HF*p*EF or HF*r*EF phenotypes modified the associations in SA1 or SA2 above.

Additionally, we broadened our analysis in **SA2** to also evaluate environmental and community variables that we hypothesized could be related to BNP levels in heart failure patients, including community greenness (measured by the normalized difference vegetation index [NDVI]) and community socioeconomic deprivation (CSD). Finally, in *post hoc* analysis, we argued that the ability of the Electronic Medical Records and Genomics (eMERGE) algorithm to categorize the HF*p*EF and HF*r*EF phenotypes was based on patient factors indicative of more severe heart failure, so we repeated the analyses in SA3 relevant to SA1 to evaluate whether heart failure severity modified associations of UNGD activity with hospitalization.

7.1.3 SUMMARY OF HOSPITALIZATION STUDY (**CHAPTER 3**)

The results of our case control study of heart failure hospitalizations illustrated that 30-day metrics of UNGD activity, in various phases, were associated with increased odds of hospitalization among 9054 heart failure subjects with a total of 11,678 hospitalizations and frequency-matched control encounters. We estimated that, comparing the 4th to the 1st quartiles of UNGD activity, the pad preparation, stimulation, and production phases were associated with increased odds of hospitalization (OR [95% CI]: 1.70 [1.35 – 2.13], 1.80 [1.35 – 2.40], and 1.62 [1.07 – 2.45], respectively). We observed exposure-effect relations for the pad preparation metric (OR [95% CI]: 1.19 [1.01 – 1.40], 1.63 [1.35 – 1.97], and 1.70 [1.35 – 2.13], for the 2nd, 3rd and 4th quartiles compared to the 1st, respectively) and for the stimulation metric (OR [95 % CI]: 1.03 [0.81 – 1.31], 1.56 [1.19 – 2.04], and 1.80 [1.35 – 2.40], for the 2nd, 3rd and 4th quartiles compared to the 1st, respectively). These results were robust to increasing covariate control and a number of sensitivity analyses.

7.1.4 SUMMARY OF BNP STUDY (**CHAPTER 4**)

The results of the study of BNP revealed that 30-day metrics of UNGD activity production were associated with increased odds of a BNP laboratory measure being greater than or equal to 400 pg/mL. We observed exposure-effect relations for the production metric (OR [95% CI]: 1.19 [1.01 – 1.40], 1.63 [1.35 – 1.97], and 1.70 [1.35 – 2.13], for the 2nd, 3rd and 4th quartiles compared to the 1st, respectively). These results were robust to increasing covariate control and a number of sensitivity analyses. We did not observe significant associations between the other environmental factors (e.g., NDVI, proximity to major and minor roads, and CSD) that we evaluated with respect to elevated BNP levels.

7.1.5 Summary of **Chapter 5a**

We used categorizations of heart failure phenotype provided by Geisinger's eMERGE investigators as HF*r*EF, HF*p*EF, and not phenotyped groups (eMERGE not applied and eMERGE no phenotype). We then evaluated effect modification by phenotype on relations of UNGD activity metrics with hospitalization. We observed that HF*r*EF and HF*p*EF subjects in our study had a distribution of comorbidities that were biologically plausible and consistent with the heart failure descriptive epidemiologic literature (e.g., HF*r*EF subjects had a higher prevalence of coronary artery disease and previous myocardial infarction; HF*p*EF subjects had a higher prevalence of chronic obstructive pulmonary disease, chronic kidney disease, diabetes). This supported the validity of the eMERGE phenotyping algorithm and gave us confidence that phenotypes were accurately identified within the hospitalization study population. We found that HF*r*EF subjects had higher odds of hospitalization than HF*p*EF subjects, independent of UNGD activity. We also found that the not phenotyped groups had reduced odds of hospitalization compared to the HF*p*EF subjects, independent of UNGD activity. After adjusting for phenotype indicators, we evaluated UNGD activity with odds of hospitalization, and we found that phenotype could have partially confounded associations presented in **Chapter 3**, as associations of UNGD activity with hospitalization were slightly attenuated but still present for the pad preparation, stimulation, and production metrics. After evaluating cross-products between each phenotype and each respective UNGD metric in our adjusted models, we found no evidence consistent with our *a priori* hypothesis that the associations of UNGD metrics with hospitalization were stronger among those with HF*p*EF (vs HF*r*EF). We did, however, find that HF*r*EF phenotype, compared to HF*p*EF phenotype, modified associations between pad preparation metric and hospitalization. Our *post hoc* analysis

evaluated effect modification of the associations between UNGD activity metrics and hospitalization by an indicator of severity, which we generated as a binary indicator for being a HF*p*EF or HF*r*EF subject. We generated this indicator because we observed that HF*p*EF and HF*r*EF subjects in the hospitalization analysis had a shorter duration of disease, higher proportion of deceased subjects at the end of the study period, and a greater number of comorbidities. In this *post hoc* analysis, we observed that after including an indicator for severity and cross-products between this indicator and respective UNGD metric quartiles, this severity indicator modified associations between the pad preparation and spud metrics and hospitalization.

7.1.6 Summary of **Chapter 5b**

Similar to the results of **Chapter 5a**, we used eMERGE assignments to categorize subjects in our BNP study (**Chapter 4**) by phenotype category (i.e., HF*r*EF, HF*p*EF, and not phenotyped [including eMERGE not applied and eMERGE no phenotype groups]). Descriptive differences between phenotype groups were similar in the BNP study population as they were within the case-control study population, although we found that the duration of heart failure did not differ between the phenotype groups, as it did for subjects in **Chapter 5a**. We found that HF*r*EF subjects had higher odds of BNP greater than or equal to 400 pg/mL, independent of UNGD activity. However, after adjusting for indicators of phenotype groups in the models of BNP, we observed exposure-effect relations among HF*p*EF subjects with the production metric (OR [95% CI]: 1.36 [1.08 – 1.73], 1.45 [1.06 – 1.97], and 1.54 [1.08 – 2.21], for the 2nd, 3rd and 4th quartiles compared to the 1st, respectively). Unlike in **Chapter 5a**, we did not have an *a priori* hypothesis regarding which phenotype group would have stronger associations with UNGD activity and BNP levels. We found some evidence of effect

modification by HF*r*EF vs. HF*p*EF phenotype groups, but the associations between UNGD activity metrics and BNP among the HF*r*EF subjects were non-monotonic. Therefore, we did not want to over-interpret these results. Additional studies should explore differences in associations between environmental factors and BNP by phenotype.

7.1.7 Overall summary of dissertation findings

Because we evaluated associations of UNGD activity metrics with two heart failure outcomes (hospitalization and BNP), and we evaluated effect modification by eMERGE-designated heart failure phenotypes of these two primary associations (in both *a priori* and *post hoc* combinations of phenotypes), we summarized the associations observed in **Chapters 3**, **4**, **5a**, and **5b** in **Table 7.1**. In the primary evaluations of UNGD activity metrics with hospitalization (**Chapter 3**) and BNP levels (**Chapter 4**), we hypothesized that we would observe exposure-effect associations with increasing quartiles of UNGD activity and increasing odds of the respective outcome. In **Chapter 5a**, we observed that HF*r*EF subjects had higher odds of hospitalization compared to HF*p*EF subjects, independent of UNGD activity. Our *post hoc* analysis evaluated an indicator of heart failure severity (i.e., having either HF*p*EF or HF*r*EF phenotype vs. no phenotype) and found that heart failure severity modified the association of UNGD activity metrics and hospitalization. In **Chapter 5b**, we observed a stronger independent association of the HF*r*EF phenotype indicator with odds of BNP $\geq$ 400 pg/mL. Although we observed globally significant cross-products between UNGD activity metrics and HF*r*EF phenotype indicators, we could not make any conclusion regarding effect modification by HF*r*EF phenotype on the association of UGND activity and BNP because exposure-effect associations between UNGD activity metrics and BNP among the

HF*r*EF subjects appeared to be non-monotonic, and therefore difficult to interpret. Because we observed differential associations between UNGD activity and BNP levels by HF*r*EF and HF*p*EF phenotypes, we did not evaluate effect modification of this association by an indicator of severity, as we did in *post hoc* analysis in **Chapter 5a**.

Table 7.1 Summary of UNGD associations with heart failure outcomes

Outcome	UNGD Metric	Chapters 3 & 4 Main effect associations	Chapters 5a & 5b Effect modification by different combinations of phenotype groups		
			HF*p*EF (vs. 1st quartile HF*p*EF)	HF*r*EF (vs. 1st quartile HF*p*EF)	Severity indicator (HF*p*EF + HF*r*EF vs. no phenotype)
Heart failure hospitalization	Pad	Exposure-effect	4th Q association	3rd & 4th Q associations	Exposure-effect
	Spud	Null	3rd Q association	4th Q association	4th Q association
	Stim	Exposure-effect	Exposure-effect	Exposure-effect	Exposure-effect
	Prod	Exposure-effect	4th Q association	3rd & 4th Q associations	Exposure-effect
BNP (high vs. low divided at 400 pg/mL)	Pad	Null	4th Q association	No exposure-effect	
	Spud	Null	Null	No exposure-effect	
	Stim	Null	Null	Inverted U-shape*	
	Prod	Exposure-effect	Exposure-effect	Inverted U-shape*	

Abbreviations: BNP = B-type natriuretic peptide; HF*p*EF = heart failure with preserved ejection fraction; HF*r*EF = heart failure with reduced ejection fraction; Q = quartile; prod = production metric; stim = stimulation metric.
* Shape of these associations should be interpreted with caution due to overlapping confidence intervals

In **Chapters 3** & **4**, in our evaluation of the associations between UNGD activity metrics and heart failure hospitalization and BNP levels, we observed evidence that supported our *a priori* hypotheses that metrics of UNGD activity would be associated with increased odds of hospitalization and increased odds of BNP levels ≥ 400 pg/mL. However, as **Table 7.1** summarizes, these associations were not observed for all metrics of UNGD activity. In the hospitalization analysis (**Chapter 3**), we observed exposure-effect associations with the pad preparation, stimulation, and production metrics and increased odds of hospitalization. Associations with quartiles of the spud metric and hospitalization were null. In the BNP analysis (**Chapter 4**), we observed null associations for the pad preparation, spud, and stimulation metrics with odds of a BNP level ≥ 400 pg/mL; however, we observed exposure-effect associations with increasing quartiles of UNGD production. Our *a priori* hypothesis in **Chapter 5a**, which evaluated effect modification by heart failure phenotypes on the associations between UNGD

activity and hospitalization, was that we would observe stronger associations with UNGD activity among HF*p*EF (vs. HF*r*EF) subjects. Our rationale for this hypothesis was due to evidence that HF*p*EF subjects often have a higher prevalence of comorbidities (e.g., hypertension, obesity, diabetes) that are affected by environmental conditions compared to HF*r*EF subjects [1]. We did not find evidence to support this hypothesis. Instead, we observed similar odds of hospitalization for heart failure in relation to UNGD activity metrics for HF*p*EF and HF*r*EF subjects. Given these findings and our observations regarding the distribution of comorbidities and medications across HF*p*EF, HF*r*EF, and not phenotyped subjects, we evaluated, in *post hoc* analyses, effect modification of the associations between UNGD activity metrics and hospitalization by an indicator of severity. We did observe that this indicator of severity modified associations between the pad preparation and spud metrics and hospitalization, and after accounting for cross-products between this indicator and UNGD activity, we observed a 4th quartile association with the spud metric, which was not observed in **Chapter 3**. Lastly, in **Chapter 5b**, we did not have *a priori* hypotheses regarding effect modification by phenotype group on relations of UNGD with BNP because there were no prior studies that were relevant to this question, so these analyses were largely exploratory. As **Table 7.1** outlines, we did observe associations between UNGD activity and BNP among HF*p*EF subjects for the production metric, and, in contrast to what we observed in **Chapter 4**, also for the 4th quartile of the pad preparation metric. Associations between UNGD metrics and BNP among HF*r*EF subjects, who had greater odds of BNP $\geq$ 400 pg/mL independent of UNGD activity, evidenced no clear exposure-effect association and the non-monotonic association was frankly somewhat difficult to interpret.

7.2 COMPARISONS ACROSS STUDY POPULATIONS

Because samples of persons with the relevant data were different in the different chapters, we evaluated the degree to which bias could have accounted for differential findings across chapters. **Figure 2.1** outlines the selection process, and subsequent tables in **Chapter 2** (**Tables 2.11- 2.15**) illustrate that subjects who had the eMERGE algorithm applied had, in general, greater usage of medications and had a higher prevalence of comorbidities. We are confident that the results of **Chapters 3** and **4** are not biased by these differences, however, as we modeled the probability of being selected into either of these studies (from the 13,183 subjects eligible for selection) and then we included sensitivity analyses that accounted for inverse probability weights in both **Chapters 3** and **4**. The results of inverse probability weighted sensitivity analyses indicated that neither associations nor inferences substantially differed in the unweighted and weighted analyses.

One notable difference across study populations, however, is that subjects in **Chapters 4** and **5b** appeared to have more severe disease than those included in **Chapters 3** and **5a**. For example, a greater proportion of subjects in **Chapters 4** and **5b** were deceased by the end of the study period compared to the proportion of subjects who were deceased by the end of the study period in **Chapters 3** and **5a** (48.8% vs. 33.7%). Although the two samples did not differ greatly in terms of the duration of heart failure, sex, age, race/ethnicity, or community type, the mean (SD) number of medication orders for heart failure among subjects in the BNP analysis was 62.8 (131) over the duration of the study period compared to a mean of 38.6 (96.1) for heart failure among subjects in the hospitalization analysis (**Chapter 2, Table 2.12**). This is reasonable considering that monitoring and measurement of BNP would be more common among more severe heart failure subjects.

7.3 UNGD AND HEART FAILURE: CONTRIBUTION TO THE EPIDEMIOLOGIC LITERATURE

Heart failure is a common, severe disease that primarily affects older adults, with approximately 5.7 million persons living with heart failure in the United States [2, 3]. Heart failure is the most common diagnosis among hospitalizations for elderly individuals [4], and these hospitalizations carry a substantial monetary cost [5]. Given the prevalence and public health burden of this disease, it was important to understand the environmental contributions to heart failure, which has been shown to be exacerbated by environmental conditions [6]. A growing number of epidemiologic studies have documented associations between UNGD activity and low birth weight and small for gestational age [7, 8]; preterm birth [9-11]; congenital defects [12]; three types of asthma exacerbations [13, 14]; migraine, fatigue, and nasal and sinus symptoms [15]; and depression symptoms [16]. However, no studies, to date have evaluated associations between UNGD activity and any disease that primarily affects older individuals. This dissertation research has been the first to evaluate associations between UNGD activity and heart failure outcomes.

Further, this is the first study to evaluate UNGD activity in relation to any biological marker. This is an important contribution to both the epidemiologic literature on UNGD associations with health outcomes, as well as to the environmental epidemiologic literature of BNP, which has only been studied in relation to environmental factors in very small sample sizes with mixed results [17-19]. The findings of associations between UNGD production activity with BNP (**Chapter 4**) are biologically plausible given how we hypothesized air pollution and stress (i.e., the two main pathways through which UNGD impacts cardiac health) would be associated with heart failure outcomes. These findings have critical public health relevance, since BNP levels have been associated with mortality in several epidemiologic studies [20-22].

Lastly, this study is the first to have evaluated any environmental associations with heart failure subjects phenotyped by the eMERGE heart failure phenotyping algorithm [23, 24] and the first to our knowledge to evaluate effect modification by important disease categories of relations of UNGD with any health outcome. This is a novel contribution to heart failure epidemiology, as studies of heart failure outcomes have been limited by the inability to systematically differentiate heart failure diagnoses obtained from electronic health records by phenotype groups [25, 26]. Although we did not observe differences in associations between UNGD activity and hospitalizations by subjects with HF*r*EF or HF*p*EF, we did find that severity of disease modified the associations we observed in **Chapter 3**, which supports the validity of our primary findings and suggests that more severe subjects had stronger associations between UNGD activity metrics and the likelihood of hospitalization, which is a biologically plausible finding.

7.4 CAUSAL INFERENCES BASED ON CURRENT BODY OF EVIDENCE

What is clear regarding the exposure scenario of UNGD is that, although we have not implicated a specific agent or pathway in the observed associations between UNGD activity and heart failure outcomes, there is enough consistent evidence now to conclude that UNGD is associated with negative impacts on population health [7, 9, 11, 13, 15, 16, 27-30]. Epidemiologic studies have found associations between UNGD activity metrics and several health outcomes. The main advantage of using UNGD activity metrics to quantify potential exposures is that these can be assessed retrospectively; the main disadvantage of this approach is that studies of UNGD activity metrics and health outcomes are vulnerable to spatial and temporal confounding. In this study and in several studies of UNGD activity and health outcomes within the Geisinger

population, the extent to which unmeasured spatial and temporal confounding influenced results was assessed. In our **Chapter 3** analysis, we conducted a negative exposure control analysis, assigning UNGD activity metrics from 2014 and 2015 to hospitalizations and matched control dates in 2008 and 2009, and we found null associations with hospitalization. The production metric had an elevated, but not statistically significant, association with hospitalization, but the production metric was also highly correlated over time. Other studies in this population have performed negative outcome control analyses (i.e., evaluating UNGD activity in relation to a biologically implausible outcome), and all have found null associations, for diarrheal illness [13], skin and soft tissue infections [11], and cold/flu, ear pain, or bad breath symptoms [15]. These sensitivity analyses were constructed to test if causal assumptions [31, 32], particularly temporality (negative exposure control) and biologic plausibility (negative exposure control analyses), were violated in assessment of UNGD activity and health outcomes, and they were not. The epidemiologic evidence of associations between UNGD activity and adverse health outcomes is also growing, and findings have been consistent across studies in Colorado [28], Texas [9, 33], and Pennsylvania [7, 11, 27], adding to the support for causal inference.

7.5 Policy implications

Due to environmental and health concerns of UNGD, some states with undeveloped shale reserves have banned their development. These bans have been in place since 2012 in Vermont, 2014 in New York state, and since 2017 in Maryland. Rationale for these bans cite risks of negative impacts to the environment and to health [34]. Interestingly, epidemiologic evidence of negative health impacts associated with UNGD has been generated by studies in Pennsylvania. However, Pennsylvania has not banned hydraulic fracturing and development of shale gas resources. This is likely

because there are substantial short-term economic benefits to UNGD, and Pennsylvania is the second-largest natural gas producer in the United States (after Texas) [35]. However, economic benefits must be viewed through a holistic lens that internalizes the costs for hospitalizations and overall disease burdens likely attributable to UNGD.

Perhaps most concerning is that UNGD is a relatively recent phenomenon, and what we know about the population health impacts is limited to associations between UNGD activity and short-term health impacts (i.e., hospitalizations, birth outcomes, acute symptoms). Given the short-term impacts that have been observed, it is possible that long-term health impacts will be realized in the future. We also do not know the cumulative impact on population health, which is especially worrisome because UNGD production is projected to increase over the next 30 years [36].

The probable health effects associated with UNGD activity should be considered not just from an environmental standpoint (i.e., regarding policies that allow for shale gas development), but also from a health services and health policy standpoint, since it is clear that environmental factors contribute to the burden of cardiovascular disease [37-40]. This has been evident well before this dissertation research linked UNGD activity to hospitalizations for heart failure and to increased levels of BNP. Health care systems, insurers, and clinicians should therefore be more vocal about environmental activities that contribute to the burden of cardiovascular disease, especially regarding hospitalizations for heart failure, which are costly [41]. This research has demonstrated that UNGD activity was associated with two different heart failure outcomes, one of which was a mechanistically important biomarker that itself predicts life expectancy, and these associations were modified by heart failure phenotypes in biologically plausible ways. These findings would seem to strengthen the causal evidence in our view.

7.5 Final remarks

Heart failure is a prevalent disease with a high mortality rate and public health burden. Patients living with heart failure are limited in terms of mobility and activity, and they are at risk for frequent hospitalization. We observed biologically plausible and robust associations with metrics of UNGD activity and hospitalization and with greater odds of an elevated BNP level, a biological indicator of diagnosis and prognosis, among heart failure patients. We also observed that heart failure severity modified the associations between UNGD activity and hospitalization for heart failure, suggesting that UNGD activity in Pennsylvania has contributed to the public health burden of heart failure in this population.

7.6 References

1. McHugh, K., et al., *Heart Failure With Preserved Ejection Fraction and Diabetes.* Journal of the American College of Cardiology, 2019. **73**(5): p. 602.
2. Ziaeian, B. and G.C. Fonarow, *Epidemiology and aetiology of heart failure.* Nature Reviews Cardiology, 2016. **13**: p. 368.
3. Mozaffarian, D., et al., *Heart Disease and Stroke Statistics—2016 Update: A Report From the American Heart Association.* Circulation, 2016. **133**(4): p. e38-e360.
4. Fang, J., et al., *Heart Failure-Related Hospitalization in the U.S., 1979 to 2004.* Journal of the American College of Cardiology, 2008. **52**(6): p. 428-434.
5. Heidenreich, P.A., et al., *Forecasting the Impact of Heart Failure in the United States: A Policy Statement From the American Heart Association.* Circulation: Heart Failure, 2013.
6. Shah, A.S.V., et al., *Global association of air pollution and heart failure: a systematic review and meta-analysis.* The Lancet. **382**(9897): p. 1039-1048.
7. Stacy, S., et al., *Perinatal outcomes and unconventional natural gas operations in Southwest Pennsylvania.* PLoS One, 2015. **10**(6): p. e0126425.
8. Currie, J., M. Greenstone, and K. Meckel, *Hydraulic fracturing and infant health: New evidence from Pennsylvania.* Science Advances, 2017. **3**(12): p. e1603021.
9. Whitworth, K.W., A.K. Marshall, and E. Symanski, *Maternal residential proximity to unconventional gas development and perinatal outcomes among a diverse urban population in Texas.* PloS one, 2017. **12**(7): p. e0180966-e0180966.
10. Walker Whitworth, K., A. Kaye Marshall, and E. Symanski, *Drilling and Production Activity Related to Unconventional Gas Development and Severity of Preterm Birth.* 2018(1552-9924 (Electronic)).
11. Casey, J.A., et al., *Unconventional Natural Gas Development and Birth Outcomes in Pennsylvania, USA.* Epidemiology, 2015.
12. Janitz, A.E., et al., *The association between natural gas well activity and specific congenital anomalies in Oklahoma, 1997–2009.* Environment International, 2019. **122**: p. 381-388.
13. Rasmussen, S.G., et al., *Association Between Unconventional Natural Gas Development in the Marcellus Shale and Asthma Exacerbations. .* JAMA Internal Medicine, 2016(2168-6114).
14. Willis, M.D., et al., *Unconventional natural gas development and pediatric asthma hospitalizations in Pennsylvania.* (1096-0953 (Electronic)).
15. Tustin, A.W., et al., *Associations between Unconventional Natural Gas Development and Nasal and Sinus, Migraine Headache, and Fatigue Symptoms in Pennsylvania.* Environ Health Perspect, 2016.
16. Casey, J.A., et al., *Associations of unconventional natural gas development with depression symptoms and disordered sleep in Pennsylvania.* Scientific Reports, 2018. **8**(1): p. 11375.
17. Wellenius, G.A., et al., *Effects of ambient air pollution on functional status in patients with chronic congestive heart failure: a repeated-measures study.* (1476-069X (Electronic)).
18. Emiroglu, Y., et al., *BNP levels in patients with long-term exposure to biomass fuel and its relation to right ventricular function.* (1522-9629 (Electronic)).
19. Vieira, J.L., et al., *Respiratory Filter Reduces the Cardiovascular Effects Associated With Diesel Exhaust Exposure: A Randomized, Prospective, Double-*

Blind, Controlled Study of Heart Failure: The FILTER-HF Trial. (2213-1787 (Electronic)).
20. York, M.K., et al., *B-Type Natriuretic Peptide Levels and Mortality in Patients With and Without Heart Failure.* Journal of the American College of Cardiology, 2018. **71**(19): p. 2079-2088.
21. Vodovar, N. and D. Logeart, *Similar BNP and Mortality Association in Patients With and Without Heart Failure.* Journal of the American College of Cardiology, 2018. **71**(19): p. 2089.
22. Santaguida, P.L., et al., *BNP and NT-proBNP as prognostic markers in persons with acute decompensated heart failure: a systematic review.* (1573-7322 (Electronic)).
23. Bielinski, S.J., et al., *A Robust e-Epidemiology Tool in Phenotyping Heart Failure with Differentiation for Preserved and Reduced Ejection Fraction: the Electronic Medical Records and Genomics (eMERGE) Network.* 2015(1937-5395 (Electronic)).
24. Bielinski;, S.J., *Heart Failure (HF) with Differentiation between Preserved and Reduced Ejection Fraction*, in *PheKB*, M. Clinic., Editor. 2013.
25. Conrad, N., et al., *Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals.* The Lancet, 2018. **391**(10120): p. 572-580.
26. Blecker, S., et al., *Trends in Hospital Readmission of Medicare-Covered Patients With Heart Failure.* Journal of the American College of Cardiology, 2019. **73**(9): p. 1004-1012.
27. Hill, E.L., *Shale gas development and infant health: Evidence from Pennsylvania.* Journal of Health Economics, 2018. **61**: p. 134-150.
28. McKenzie, L.M., et al., *Birth outcomes and maternal residential proximity to natural gas development in rural Colorado.* Environmental health perspectives, 2014. **122**(4): p. 412-417.
29. Rabinowitz, P.M., et al., *Proximity to natural gas wells and reported health status: results of a household survey in Washington County, Pennsylvania.* Environ Health Perspect, 2015. **123**(1): p. 21-6.
30. Werner, A.K., et al., *Environmental health impacts of unconventional natural gas development: A review of the current strength of evidence.* Science of The Total Environment, 2015. **505**: p. 1127-1141.
31. Fedak, K.M., et al., *Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology.* Emerging themes in epidemiology, 2015. **12**: p. 14-14.
32. Susser, M., *What is a Cause and How Do We Know One? A Grammar for Pragmatic Epidemiology.* American Journal of Epidemiology, 1991. **133**(7): p. 635-648.
33. Walker, W.K., K.A. Marshall, and E. Symanski, *Drilling and Production Activity Related to Unconventional Gas Development and Severity of Preterm Birth.* Environ Health Perspect. , 2018. **126**(3): p. 037006.
34. Health;, N.Y.S.D.o., *A public health review of high volume hydraulic fracturing for shale gas development*. 2014.
35. EIA;, U., *Pennsylvania State Energy Profile.* 2018.
36. EIA;, U., *Annual Energy Outlook 2019 with projections to 2050*. 2019, US EIA.
37. Bhatnagar, A., *Environmental Cardiology.* Circulation Research, 2006. **99**(7): p. 692.
38. Giorgini, P., et al., *Air Pollution Exposure and Blood Pressure: An Updated Review of the Literature.* (1873-4286 (Electronic)).

39. Rajagopalan, S., S.G. Al-Kindi, and R.D. Brook, *Air Pollution and Cardiovascular Disease: JACC State-of-the-Art Review.* (1558-3597 (Electronic)).
40. Bourdrel, T., et al., *Cardiovascular effects of air pollution.* (1875-2128 (Electronic)).
41. Kilgore, M., et al., *Economic burden of hospitalizations of Medicare beneficiaries with heart failure.* Risk management and healthcare policy, 2017. **10**: p. 63-70.

Appendix A. Heart failure algorithm

Heart Failure

With Differentiation between Reduced and Preserved Ejection Fraction

Phenotype Algorithm Pseudo Code

Mayo Clinic

Version: March 4, 2014

Contacts:
Suzette J. Bielinski (Bielinski.suzette@mayo.edu, 507-538-4914)
Jyotishman Pathak (Pathak.jyotishman@mayo.edu, 507-538-8384)

Updates:

- March 4, 2014:
  - Updated pseudo code to include instructions for categorizing free test ejection fraction results.
  - Data Dictionary was altered to include optional reporting of free text EF results when numeric EF measurements are unavailable.
  - Height and weight units were changed to cm and kg in the data dictionary to correspond to eMERGE standard practice.

Introduction:

Heart failure (HF) is a complex syndrome characterized by the inability of the heart to supply sufficient blood flow to the body. HF is diagnosed clinically and further dichotomized by left ventricular ejection fraction (i.e. reduced or preserved). In 2010, HF affected 6.6 million Americans at a cost of 34.4 billion [279, 280]. However, the syndromic nature of HF presents challenges in identification of HF cases and controls from EHR data for research given that the diagnosis is clinical. The Electronic Medical Records and Genomics (eMERGE) Network[281] consortium has demonstrated the applicability and portability of EHR derived phenotype algorithms using different types and modalities of clinical data for algorithm execution including billing and diagnoses codes, natural language processing (NLP), laboratory measurements, patient procedure encounters, and medication data.

Development:

Using a gold standard cohort of 706 manually abstracted HF cases defined according to Framingham Heart Failure Criteria[282] from the Heart Failure in the Community Cohort (R01 HL72435), structured EHR data were combined with analyses of the clinical note (unstructured) to identify the set of parameters needed to reidentify all the cases.

HF terms were identified using natural language processing (NLP) (i.e. dictionary lookup, negation/probable identification with ConText[283-285]) to identify positive hits of HF from the major and secondary problem lists of the clinical note.

Algorithm and Covariates:

This algorithm requires the following types of information:

- Patient demographics
- Encounter history
- ICD diagnosis codes
- Structured problem list or unstructured problem list (processed using NLP)
- Echocardiography measurements
- Medications

Some covariates are repeated measurements requiring associated dates and others are a single measure.

Collection of Covariates

Detailed definitions of all covariates are included in the data dictionary. Note that some covariates are captured as repeated measures and include ejection fraction, myocardial infarction, BMI, and medication use. Medication history includes the following list of drugs. A complete list of drug names is included in the data dictionary.

- Angiotensin Converting Enzyme Inhibitors (ACE)
- Angiotensin II Receptor Blockers (ARB)
- Beta Blockers (BB)
- Calcium Channel Blockers (CCB)

Implementation Notes:

- Appendix A Mayo Clinic Implementation: provides a description of how the algorithm was implementation a CDA formatted document structured EMR with structured echo data.
- Appendix B Group Health Implementation: provides a description of how the algorithm was implementation a non-CDA formatted EMR with unstructured echo data.
- Appendix C Alternative Case Definition: provides a description of an optional NLP only case definition. Consider at institutions with highly transient patient populations.

For Mayo and Group Health, the NLP component of the HF algorithm was implemented using MedTagger-IE, a pattern-based information extraction framework, which will be available open source under Open Health Natural Language Processing (OHNLP) consortium (Liu H**, Bielinski SJ**, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, Ravikumar KE, Wu ST, Kullo IJ, Chute CG. An information extraction framework for cohort identification using electronic health records. AMIA Summits Transl Sci Proc. 2013;2013:149-153. PMCID: PMC3845757).

Step 1: Identifying Heart Failure Case Definition (both conditions must be met)

Presence of ICD9-CM Diagnosis Codes for HF

- Primary Heart Failure Codes = (428.X)

AND

Positive mention of HF in the problem list through either NLP or structured problem list

- Unstructured problem list (NLP) – at least one positive mention of a HF term in diagnosis-related sections. Positive mention is defined using ConText for assigning statuses to each NLP result – positive, probable, and negative (Chapman, Chu et al. 2007; Harkema, Dowling et al. 2009; Chapman, Lee et al. 2011). Thus a positive hit for this requirement equates to a non-negative and non-probable result. Mapping of terms is insensitive to upper/lower case.
    - multi-organ failure or multiorgan failure
    - cardiac failure
    - CHF
    - heart failure
    - ventricular failure
- Structured problem list – The descendant traversal of SNOMEDCT code 84114007 (heart failure).

Step 2: Date of First Documented Heart Failure

Data Processing for Problem List

- Unstructured problem list (NLP) **-** assign the earliest note date among those notes with non-negative and non-probable HF terms detected in diagnosis-related sections.
- Structured problem list **-** assign the earliest date among the corresponding HF codes appearing in the problem list.

Case Date Assignment

- If the earliest problem list date and earliest ICD9 date fall within 1 year of each other assign the patient the earlier of the two dates.
- If the earliest problem list date and earliest ICD9 date are assigned > 1 year apart then do as follows
    - Use the first two dates that occur within a rolling year between any combination of problem list dates and ICD9 dates. If two dates found within a rolling year, use the earlier one.
    - Exclude cases where a date could not be assigned because they did not have any ICD9 nor problem list dates within a rolling year of each

other (Note: If the same date appeared twice within the problem list dates, this was ignored).

Step 3: Classifying Heart Failure in terms of Ejection Fraction

Classify the type of HF using the priority metric below. Some patients may have multiple echo measurements within a given time window – use the lowest EF recorded in the time window.

Numeric EF Results - Priority Metric

1. Lowest EF measured 0-182 days (approximately a 6 month period) **after** the HF date. If missing, go to number 2.
2. Lowest EF measured 0-182 days (approximately a 6 month period) **prior** to the HF date. If missing go to number 3.
3. Lowest EF measured 183-365 days **after** the HF date. If missing go to number 4.
4. Lowest EF measured 183-365 days **prior** to HF date. If missing, set HF type to none (HF Type = 0).

Coding Rules to Assign Heart Failure Type:

- HF with reduced EF = ejection fraction < 50% (HF Type = 1)
- HF with preserved EF = ejection fraction ≥ 50% (HF Type = 2)
- No qualifying EF measurements within any of the time frames considered(HF Type = 0)

Free Text Ejection Fraction Results - Priority Metric
(Optional – Use only if numeric EF measurements are not available)

EF Result Categories	Free Text Variations
Preserved	normal, supernormal, low-normal, moderate
Reduced	abnormal, reduced, low, severe, decreased

1. EF Result 0-182 days **after** the HF date
    a. If EF preserved then HF Type = 2
    b. If EF reduced then HF Type = 1
    c. If missing, go to number 2.
2. EF Result 0-182 days **prior** the HF date
    a. If EF preserved then HF Type = 2
    b. If EF reduced then HF Type = 1
    c. If missing, go to number 3.
3. EF Result 183-365 days **after** the HF date
    a. If EF preserved then HF Type = 2
    b. If EF reduced then HF Type = 1
    c. If missing, go to number 4.
4. EF Result 183-365 days **prior** the HF date
    a. If EF preserved then HF Type = 2
    b. If EF reduced then HF Type = 1
    c. If missing, set HF type to none (HF Type = 0).

Step 4: Identifying Heart Failure Controls (all conditions must be met):

- Absence of any heart failure ICD9 codes in medical record

AND

- Absence of positive heart failure terms in the EMR via NLP at any section or absence of heart failure codes in the structured problem lists

AND

- EF ≥ 50% if measured or patient does not have echocardiographic measurements

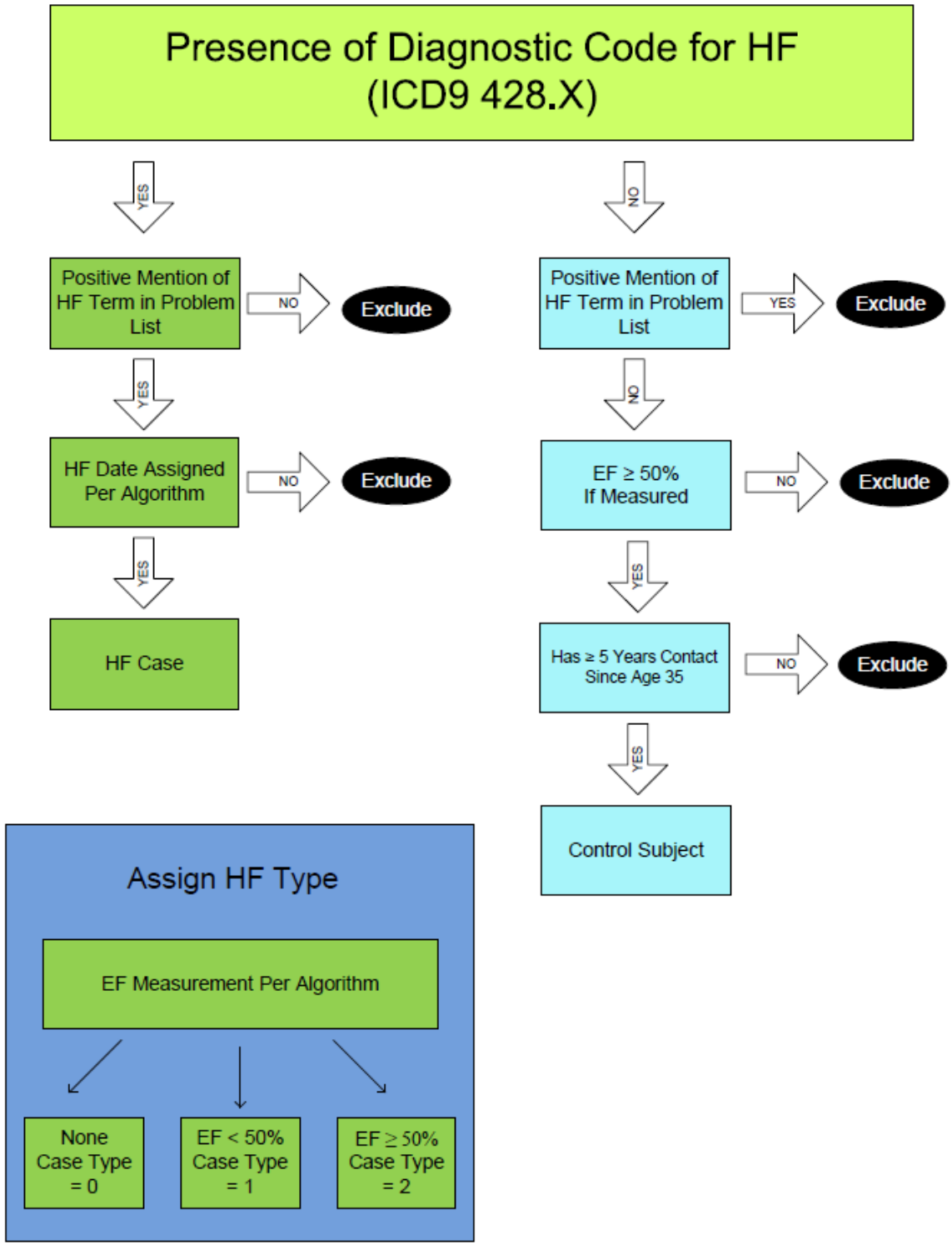
AND

- Has ≥5 years of continuous enrollment/contact since age 35.

    Note: This is the same definition used in the Zoster algorithm. "Continuous enrollment/contact" is implemented in an integrated group practice (HMO) setting as a period of continuous enrollment ignoring gaps of up to 90 days (which tend to reflect administrative data lags rather than actual interruptions in access to care). In fee-for-service or non-integrated care settings, continuous

enrollment is implemented as consecutive 5-year periods with at least one encounter per period. At Mayo Clinic we used ICD9 and CPT dates to define encounters. Encounters may be rolled up to the calendar date for purposes of establishing encounters during a period.

Heart Failure Algorithm Flow Chart



Appendix A: Mayo Clinic Implementation:

NLP Specific Details: The NLP component of the case definition was implemented by searching the Major and Secondary problem list section of the clinical note for at least one positive mention of one of the heart failure terms. Positive mention is defined using ConText for assigning statuses to each NLP result – positive, probable, and negative [283-285]. Thus a positive hit for this requirement equates to a non-negative and non-probable result. Mapping of terms is insensitive to upper/lower case. All NLP dates associated with probable heart failure as defined by ConText were excluded. Among the remaining dates, we assigned the earliest NLP date among those associated with the major problem list and in the case where there was no note date associated with the major problem list; the earliest NLP date among those associated with the secondary problem list was used.

Echocardiography Results: Mayo Clinic echo data is stored in a structured database. Variable corresponding to ejection fraction (EF) measurements were identified. It is common for several EF measurements to be taken during a single exam thus the average of all available EF measurements from a single exam was used.

Appendix B: Group Health Implementation:

NLP Specific Details: The clinical notes are in non-CDA formatted documents, thus SecTag was used to detect Diagnosis and other sections (i.e. Chief Complaints or Impressions as the Secondary Problem List section).

Echocardiography Results: Group Health echo data is unstructured thus NLP was deployed to search the radiology reports for EF measurements. The following list of *regular expressions for reporting EF was used.

- Calculated EF ##%
- Calculated LVEF ##%
- Calculated LV ejection fraction ##%
- Calculated Left Ventricular ejection fraction ##%
- Calculated Ejection Fraction ##%
- Calculated Ejection Fraction ##%. Visual estimate ##%-##%
- Estimated EF ##%
- Estimated EF = ##%
- Estimated EF ##%-##%
- Estimated Ejection Fraction ##%
- Estimated Ejection Fraction ##%-##%
- Estimated Left Ventricular Ejection Fraction ##%
- Estimated Left Ventricular Ejection Fraction ##%-##%
- Estimated Left Ventricular Ejection Fraction range ##%-##%
- EF ##%
- Ejection Fraction ##%
- LVEF ##%
- LVEF ~ ## - ##%Left Ventricular Ejection Fraction ##%
- Visual Estimate of LVEF ##%
- Visual estimate of Left Ventricular Ejection Fraction ##%
- Visual Estimate of EF ##%
- Visual Estimate of Ejection Fraction ##%

*The regular expression list above includes the variations identified at Mayo Clinic and Group Health and thus is not an exhaustive list of every possible combination of the use of characters such as "=" or "~".

Appendix C: Alternative Case Definition:

As part of the algorithm development, an alternative case definition using NLP evidence in the absence of an ICD9 diagnosis HF code (i.e. 428.X) was tested. This alternative case definition at Mayo Clinic required a positive NLP hit for HF terms 1 time in the Major problem list AND ≥ 5 times in Secondary problem list. HF date assignment considered only the NLP dates and those patients without two dates in a rolling year are excluded. Despite good overall performance), the yield of cases was minimal (n = 13 at Mayo Clinic). Abstraction of these cases revealed that all of the 13 patients identified by this definition were true cases and all were transient/referral patients. Therefore the NLP only case definition was not included in the final algorithm. However, for sites with a large referral or transient population, the alternative case definition should be considered as it may result in higher yield as compared to the stable population at Mayo Clinic.

Appendix B. Documentation of IRB approval

Geisinger Institutional Review Board (GIRB)
Geisinger Medical Center
100 N. Academy Avenue
Danville, PA 17822-3069
570-271-8663Tel

Geisinger

Approval Notice – Expedited
Continuing Review

October 03, 2018

Brian S Schwartz, MD, MS
GMC - Center for Health Research

IRB #: **2013-0114 (Marcellus Health Prospective)**, entitled ***Marcellus Shale and Health Prospective Analysis***

RE: Geisinger Continuing Review Form, 10/02/2018 07:56:13 AM EDT

Dear Brian S Schwartz, MD, MS:

The continuing review for the above study was reviewed and approved via expedited review on 10/03/2018.

Please note the following information about your IRB approval:

Submission Components		
Form Name	Version	Outcome
Geisinger Continuing Review Form	Version 6.0	Approve

Geisinger IRB "2018 New Rule" Transition Note:
This study will **not** transition to Geisinger's 2018 New Rule requirements because it is FDA-regulated or federally-funded/sponsored. All reporting requirements, including annual continuing review, remain the same.

You are now approved to continue your study activity. The study has been approved until 10/02/2019.

Approval Period: 10/03/2018 - 10/02/2019

If you have any questions or need further help, please contact the Human Research Protection Program staff at (570) 271-8663.

Sincerely,

H. Lester Kirchner PhD
IRB Co-Chair
Institutional Review Board

William E. Crowder, Jr, MD FACOG
IRB Co-Chair
Institutional Review Board

cc: Brian S Schwartz, MD, MS, Dione G Mercer

Appendix C. Curriculum Vitae

Tara P. McAlexander, MPH, PhD Candidate

3021 N. Calvert Street, Baltimore, MD 21218 tmcalex1@jhu.edu, (856) 524-0750

EDUCATION

Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD
Doctor of Philosophy Candidate in Exposure Sciences and Environmental Epidemiology,
Department of Environmental Health and Engineering, September 2014-present.

- Dissertation: *Unconventional natural gas development (UNGD), environmental factors, and heart failure: Epidemiological associations with hospitalizations, B-type natriuretic peptide, and ejection fraction-based phenotypes*
-

Columbia University Mailman School of Public Health, New York, NY
Department of Environmental Health Sciences
Master of Public Health in Environmental and Molecular Epidemiology, May 2012

- Thesis: *Prevalence and distribution of household asthma hazards in an elderly population in Washington Heights, NYC*

Barnard College, Columbia University, New York, NY
Bachelor of Arts, May 2010
Major: Environmental Science

- Thesis: *Influence of urban street noise on neighborhood crime: A pilot field study*, passed with distinction

School for International Training, Brattleboro, VT
Study Abroad in Geneva, Switzerland: *Development Studies and Public Health*, Spring 2009

- Independent study project: *Health effects of pesticide use in Viet Nam*

PROFESSIONAL EXPERIENCE

The Fund for Public Health in New York/ New York City Department of Health and Mental Hygiene, Bureau of Environmental Surveillance and Policy
New York, NY, October 2013-August 2014, *Environmental Health Research Fellow*

- Integrated neighborhood-level air quality data from the New York City Community Air Survey (NYCCAS) with NYC hospital records in a NIH/NIEHS-funded pregnancy outcome study
- Calculated birth outcome indicators for the NYC Environmental Health Tracking Portal using the NYC Vital Records database
- Created exposure metrics and performed time-series analyses on the relationship between National Weather Service data and pregnancy outcomes using SAS and R

University of California, San Francisco School of Medicine, Department of Epidemiology and Biostatistics, Philip R. Lee Institute for Health Policy Studies
San Francisco, CA, June 2013-August 2014, *Consultant*

- Assisted with survey design, coordination of stakeholder funded engagement, and survey dissemination for a NSF-funded national study of emergency preparedness in the death care sector

Association of Schools and Programs of Public Health/U.S. Environmental Protection Agency
Research Triangle Park, NC, September 2012-August 2013, *ASPH/EPA Environmental Health Fellow*

- Developed a built and natural environmental health vulnerability index for Durham-Chapel Hill, NC at the US EPA Office of Research and Development Sustainable and Healthy Communities Research Program
- Conceptualized and executed database management and statistical plans based on geospatial and secondary datasets using SAS, R, and ArcGIS

Columbia University Mailman School of Public Health, Department of Sociomedical Sciences
New York, NY, August 2010-May 2012, *Project Coordinator*

- Managed a research team of ten graduate research assistants, project budget and timelines, coordination with community partners and programs, data integrity and security, IRB human subject protection protocols, and grant reporting requirements for a HUD-funded community-based research intervention study on environmental household hazards in a senior population of Washington Heights, New York City

Columbia University Mailman School of Public Health, Department of Sociomedical Sciences New York, NY, September 2007-January 2009; September 2009- July 2010, *Research Assistant*

University of Pennsylvania School of Medicine, Center of Excellence in Environmental Toxicology
Philadelphia, PA, Summer 2009, *Penn Undergraduate Environmental Health Scholar*

American Cancer Society, Cherry Hill, NJ Summer 2007, *Intern*

HONORS AND AWARDS

Johns Hopkins University Department of Environmental Health and Engineering Research Retreat, 1st place poster award for Exposure Sciences and Environmental Epidemiology, January 2018.

American Public Health Association Student Scholarship Recipient Sponsored by External Medical Affairs, Pfizer, Inc., 2011

American Public Health Association Environment Section's Student Poster Achievement Award, 1st place: Association between urban street noise and neighborhood crime: a pilot field study, 2010

American Public Health Association Environment Section's Student Travel Award, 2010

ABSTRACTS

McAlexander TP, Pollak J, Bandeen-Roche K, Schwartz BS. Association between unconventional natural gas development activity and hospitalization among heart failure patients in northeastern Pennsylvania. Society for Epidemiologic Research Annual Meeting; 2018, Baltimore, Maryland.

McAlexander TP, Neitzel R, Gershon RRM. Geographic and temporal variability in road/street noise levels in New York City. Conference of ISEE, ISES and ISIAQ Environmental Health Conference; 2013, Basel, Switzerland.

Bush KF, Jackson L, Sears A, **McAlexander TP**, Baynes J, Pilant A, Maxson P, Edwards SE, Miranda ML. Applying EnviroAtlas to Public Health: Investigating the association between green space and birth outcomes in Durham-Chapel Hill, NC. ESRI User Conference; 2013, San Diego, CA.

McAlexander TP, Bush KF, King K, Jackson L, Araujo R. The spatial and socio-economic distribution of healthy food options in Durham-Chapel Hill, NC. Fifth Annual Health Disparities Conference at Teachers College, Columbia University; 2013, New York, NY.

Bush KF, **McAlexander TP**, King K, Jackson L, Araujo R. Disparities in access: the spatial distribution of resources in Durham-Chapel Hill, NC. University of North Carolina Minority Health Conference; 2013, Chapel Hill, NC.

McAlexander TP, Perzanowski M, Hernández-Cordero LJ, Gershon RRM. Prevalence and distribution of environmental hazards associated with asthma in the households of elderly community members in Washington Heights, NYC. American Public Health Association Annual Meeting & Exposition; 2012, San Francisco, CA.

Gershon RRM, **McAlexander TP**, Hernández -Cordero LJ, Chu L, Chan M, Perzanowski M. Healthy homes, Healthy Seniors: Baseline Results. American Public Health Association Annual Meeting & Exposition; 2012, San Francisco, CA.

Gershon RRM, Magda LA, Riley H, **McAlexander TP**, Neitzel R. Hearing health and mass transit ridership. American Public Health Association Annual Meeting & Exposition; 2012, San Francisco, CA.

McAlexander TP, Neitzel R, Gershon RRM. The New York City urban soundscape: a pilot study. The International Conference on Urban Health; 2010, New York, NY.

McAlexander TP, Neitzel R, Gershon RRM. Association between urban street noise and neighborhood crime: a pilot field study. American Public Health Association Annual Meeting & Exposition; 2010, Denver, CO.

Gershon RRM, Pearson JM, **McAlexander TP**. A novel tool to address safety hazards in the home healthcare setting. American Public Health Association Annual Meeting & Exposition; 2010, Denver, CO.

PUBLICATIONS

Almeter A, Tashie A, Procter A, **McAlexander T**, Browning D, Rudder C, Jackson L, and Araujo R. A needs-driven, multi-objective approach to allocate urban ecosystem services from 10,000 trees. *Sustainability*. 2018, 10(12), 4488.

Savitz DA, Elston B, Bobb JF, Clougherty JE, Dominici F, Ito K, Johnson S, **McAlexander T**, Ross Z, Shmool JLC, Matte TD, Wellenius GA. Ambient fine particulate matter, nitrogen dioxide, and hypertensive disorders of pregnancy in New York City. *Epidemiology*. 2015, 26 (5): 748-757.

McAlexander TP, Gershon RRM, Neitzel RL. Street-level noise in an urban setting: assessment and

contribution to personal exposure. *Environmental Health*. 2015, 14 (18).
Gershon RRM, Sherman MF, Magda LA, Riley HEM, **McAlexander TP**, Neitzel R. Mass Transit Ridership and Self-Reported Hearing Health in an Urban Population. *Journal of Urban Health*. 2013, 90 (2): 262-75.
Neitzel RL, Gershon RRM, **McAlexander TP**, Magda LA, Pearson JM. Exposures to Transit and Other Sources of Noise Among New York City Residents. *Environmental Science and Technology*. 2012, 46 (1): 500-508.

INVITED PRESENTATIONS

Schwartz BS & **McAlexander TP**. *Shale Gas Development & Health: Research Updates*. Oral presentation at the Pennsylvania League of Women Voters conference on Shale Gas & Health, Pittsburgh, PA, November 13, 2017.
McAlexander TP. *Unconventional Natural Gas Development and Health: A Review of the Epidemiological Evidence*. Environmental Defense Fund, Washington, DC, October 20, 2016.
McAlexander TP. *Measuring neighborhood resource availability in Durham-Chapel Hill, NC.* United States Environmental Protection Agency Sustainable and Healthy Communities Research Program Seminar Series, October 29, 2013.

COMMUNITY INVOLVEMENT AND PROFESSIONAL ORGANIZATIONS

Big Brothers Big Sisters at the Y in central Maryland, *Mentor*, February 2017-present

Environmental Health and Engineering Student Organization (EHESO), Johns Hopkins Bloomberg School of Public Health, 2015-2016, *Secretary;* 2016-2017, *President-Elect*; 2017-2018, *President*

American Public Health Association, Environment Section, August 2010-2017, *Member*; December 2010-2012, *Environment Section's Liaison to the Student Assembly;* December 2010- 2015, *Co-chair, Environment Section Student Involvement Committee;* October 2015-October 2017, *Co-chair, Environment Section Program Planning Committee*

SKILLS

Software Proficiencies: PC/MAC, Word, Excel, Access, Power Point, Publisher, EndNote, HTML, ArcGIS, R statistical software, SAS, STATA, and Microsoft SQL Server.

Analytic Proficiencies: Case-control study designs, longitudinal cohort studies, time-series analyses, multi-level statistical modeling and inference, generalized estimated equations, principal components analyses, and spatial regression.